FROM TOOLS TO INSTRUMENTS: TRACKING MEASUREMENT UNCERTAINTY IN AUTONOMOUS SCIENCE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI co-scientists increasingly do work that researchers have always done themselves: they test hypotheses on their own, from planning an analysis to interpreting its results. Yet inaccurate measurements from their tools can turn a correct analysis into a wrong conclusion. Such failures are hard to identify when reference answers are unavailable. We propose a framework that treats tools as measurement instruments and carries measurement assessments through autonomous hypothesis testing. Instrument disagreement supplies an assumed error scale, and a measurement counterfactual estimates the perturbation needed to change a statistical decision. Four checks combine measurement sensitivity and instrument disagreement with sampling uncertainty, verdict–evidence contradictions, and disagreement across repeats. Their flags define a risk level computed without reference answers. We implement the framework in VERITAS-M, a multi-agent clinical co-scientist, and evaluate it on 279 hypotheses over cardiac and brain MRI, retinal images, and clinical data. VERITAS-M improves on its predecessor; using the same primary model with an additional critic, it also outperforms single-model agents in verdict accuracy. With each cohort's primary instrument, keeping conclusions flagged by at most one check lowers observed error about tenfold; the rule also lowers error for three systems it was not tuned on. Because inaccurate measurements can yield wrong conclusions that stay consistent across repeated runs, autonomous science needs what experimental science has long practiced: treating tools as instruments whose errors are measured and carried through to the conclusion.

1 INTRODUCTION

Science agents increasingly conduct research with little human involvement, from generating hypotheses to interpreting results (Boiko et al., 2023; Bran et al., 2024; Lu et al., 2026; Swanson et al., 2025; Gottweis et al., 2026). Their conclusions often depend on measurements supplied by the tools they use: tumor volumes from images, cell annotations from microscopy, or clinical variables from patient records. Errors in these measurements can change a scientific conclusion even when the statistical analysis is executed correctly.

Consider a co-scientist testing whether grade 4 gliomas contain more necrotic tissue than grade 2 gliomas. Its segmentation model reported necrosis in grade 2 tumors whose expert contours recorded none, reducing the group contrast (Fig. 1b). The analysis was reproducible, but reproducing it would preserve the error. Identifying the failure required examining the measurements.

An autonomous scientist may need to assess a conclusion before reference measurements or independent validation are available. Repeating an analysis can reveal instability, but consistent answers may still depend on inaccurate measurements. As agents produce conclusions faster than scientists can check them, each conclusion should indicate how it depends on those measurements and where further validation is needed. Autonomous workflows can make this reporting routine, attaching to each conclusion its instruments, decision sensitivity, repeat stability, and a risk level.

How can an autonomous science agent identify unreliable conclusions without access to reference answers?

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.

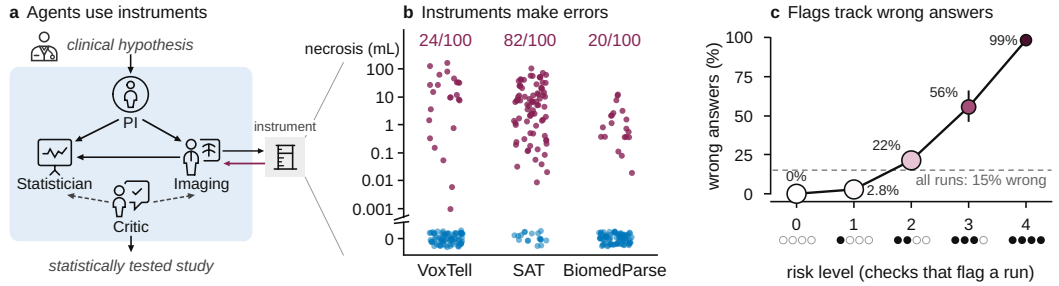

Figure 1: **Measurement errors and unreliable conclusions.** (a) An agent team tests a clinical hypothesis; the imaging agent takes its readings from a segmentation model, which we treat as an instrument. (b) Necrotic volume that three instruments report for 100 glioma patients (UCSF-PDGM) whose expert contours contain no necrosis. Red dots (counted above) are false positives; blue dots are correct zeros. The agent never sees expert contours. (c) Wrong runs (verdict or evidence label wrong) by risk level, the number of the four checks that flag a run (filled dots), over 2,308 eligible runs with each cohort's primary instrument; none of the 651 runs at level 0 is wrong. Dashed: error over all runs. Marker area: share of runs; bars: 95% bootstrap intervals over hypotheses.

We address this question with four contributions.

1. Tracking measurement uncertainty. We propose a framework that treats a measurement tool as an *instrument* and its output as a *reading*. Instrument comparisons supply an assumed error scale; a measurement counterfactual estimates the perturbation needed to change a statistical decision. Records carry these assessments from planning to interpretation (Secs. 2–3).

2. Identifying unreliable conclusions without a reference. Four measurement, sampling, and consistency checks define a risk level. With primary instruments, retaining conclusions flagged by at most one check keeps 61% of eligible runs at 1.5% observed error. The rule reduces error on three further systems; replacing instruments with less accurate ones exposes failures missed by consistency checks (Sec. 5).

3. An improved clinical co-scientist. VERITAS-M improves verdict accuracy over VERITAS (Stoffl et al., 2026) with the same models on retained benchmark items. With the same primary model as a strong single-model baseline and an additional critic, it also improves verdict accuracy on all five cohort tables; on four, the 95% interval of the gain excludes zero (Sec. 5.1).

4. A benchmark with measurement-derived reference answers. We extend the VERITAS benchmark (Stoffl et al., 2026) to 279 hypotheses across cardiac and brain MRI, retinal imaging, and clinical data. Added datasets, segmentation tools, and human graders test different measurement conditions. References use expert contours or human measurements, except for retinal quantities without expert counterparts (Sec. 4).

1.1 Relation to prior work

Agents can defer to inaccurate tools (Wang & Vemuri, 2026), and their conclusions can change with analysis choices (Rewolinski et al., 2026; Miao et al., 2026). Hypothesis-validation frameworks control statistical error under assumptions (Huang et al., 2025). Reproducible workflows make analyses traceable (Ifargan et al., 2025; Stoffl et al., 2026), while consistency checks reveal variation across answers (Raj et al., 2026; Bertran et al., 2026). Agreement across answers can persist when every run uses the same inaccurate measurements; our instrument comparisons and counterfactuals assess this dependence directly. Measurement-error methods address uncertain inputs (Carroll et al., 2006); prediction-powered inference corrects inference with predicted variables using labeled references (Angelopoulos et al., 2023). Our checks assess measurement sensitivity without runtime references. Instrument substitutions reveal wrong labels consistent with both the agent's verdict and repeated runs, yet flagged by measurement checks (Sec. 5.3; Appendix A).

2 PROBLEM FORMULATION AND RELIABILITY CHECKS

A conclusion, such as whether grade 4 gliomas contain more necrotic tissue than grade 2 gliomas, can fail because an agent misinterprets its evidence or because the evidence rests on inaccurate measurements or an incorrect analysis. These failures require different checks. We distinguish the agent's verdict from a label computed from its statistics, then assess measurement sensitivity and consistency without a reference answer.

Agent verdicts and computed evidence. Following VERITAS (Stoffl et al., 2026), a run ends in two outputs: an *evidence label* that fixed code computes from the statistics, and a *verdict* that the agents write by interpreting them. An agent specifies an analysis ϕ and cohort D. Instruments supply readings $T(D)$, and a deterministic backend computes the statistical results $S(\phi, T(D))$: the estimate, interval, p-value, and sample sizes. A fixed rule assigns the evidence label

$$L = f\big(S(\phi, T(D)), \delta\big) \in \{\textsc{Supported}, \textsc{Refuted}, \textsc{Underpowered}, \textsc{Invalid}\}, \tag{1}$$

where δ is the smallest effect size considered substantively relevant and plausibility gates are held fixed. Separately, the agent interprets the statistics to write a verdict v, one of YES, NO, or INCONCLUSIVE, with L withheld so that we can assess whether its interpretation agrees with the statistical evidence. Let m map SUPPORTED to YES, REFUTED to NO, and the remaining labels to INCONCLUSIVE. Answering NO to an underpowered null result then contradicts the run's evidence: $v \neq m(L)$.

For evaluation, L_{ref} uses the same rule with registered reference readings and analysis (Sec. 4):

Proposition 1 (Observable and latent error)**.** *A run is wrong, meaning $L \neq L_{\text{ref}}$ or $v \neq m(L_{\text{ref}})$, if and only if $L \neq L_{\text{ref}}$ or $v \neq m(L)$. The second event is a function of the run alone.*

Verdict–label contradictions are observable from the run, but consistent interpretation can preserve incorrect evidence. After removing contradictions, every remaining error involves a wrong label. Label-error evaluation therefore tests detection beyond inconsistent interpretation.

Sensitivity to measurement error. Plausibility gates invalidate readings that violate predefined constraints. For other errors, a *secondary instrument* measures the same quantity. Disagreement supplies an assumed error scale σ, conditional on error covariance and allocation between instruments (Grubbs, 1948; Joint Committee for Guides in Metrology, 2008). Assuming zero covariance, each instrument receives one third to two thirds of the disagreement variance, or its own estimate when three instruments read the same subjects; neither accounts for shared error (Appendix B).

We ask how much error would change the decision, shifting groups differently to test distortions such as the necrosis example. For a continuous biomarker, let T_d shift readings differentially across the comparison by $d\sigma$, with $T_0 = T$. For a decision function g, define

$$d^*_g = \inf\{d \geq 0 : g(S(\phi, T_d(D))) \neq g(S(\phi, T(D)))\}. \tag{2}$$

This *measurement counterfactual* expresses decision sensitivity in assumed error units, analogous to sensitivity measured in outcome events (Walsh et al., 2014) or error strength (VanderWeele & Li, 2019). With $g = f$, it targets the evidence label. The implementation searches both shift directions on a grid for g_p: significance at α and, when significant, effect direction. The full evidence label can also change when the effect crosses δ, with significance and direction unchanged. For example, a reported counterfactual of 0.5 means the search found a changed decision at half the assumed error scale along a tested direction. The first changed grid point bounds $d^*_{g_p}$ from above; it does not certify stability below that value or cover every full-label change (Appendix B). Section 5.5 tests it.

Reliability checks and the risk level. Four checks target measurement error, misinterpretation, sampling uncertainty, and variation across runs, using the run's records and two repeat executions:

- *Instrument*: the decision changes within one error unit under either error-allocation endpoint, or instruments produce different labels.
- *Label conflict*: the verdict contradicts the evidence label or the run's p-value.
- *Thin margin*: the estimate lies close to δ, the smallest relevant effect, relative to its sampling uncertainty. On δ's scale, the estimate or either interval endpoint lies less than one interval half-width from δ.

- *Repeat disagreement*: either of two repeat runs reaches a different verdict, revealing instability across executions (Bertran et al., 2026).

The number of flags defines the *risk level*:

$$r = \mathbb{1}[\text{instrument}] + \mathbb{1}[\text{label conflict}] + \mathbb{1}[\text{thin margin}] + \mathbb{1}[\text{repeat disagreement}] \in \{0, \ldots, 4\}. \quad (3)$$

Missing checks contribute zero. This ordinal score supports selection (Geifman & El-Yaniv, 2017) without assigning error probabilities. Thresholds $r < k$ are empirical: they were chosen on the design corpus and repeats are dependent. Appendix B gives the standard fixed-threshold bound.

3 A CLINICAL CO-SCIENTIST WITH MEASUREMENT CHECKS

VERITAS-M implements these checks within the multi-agent clinical workflow of VERITAS (Stoffl et al., 2026). It adds typed statistical computation to separate evidence from interpretation, and a measurement layer to carry error scales, sensitivity estimates, and instrument comparisons through the analysis (Fig. 2). Agents choose analyses and interpret results; imaging tools supply measurements; fixed code computes statistics, labels, and checks.

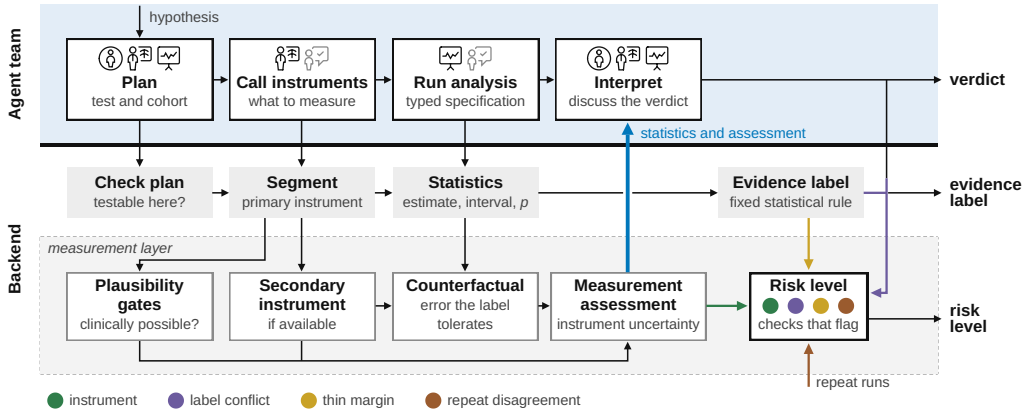

Figure 2: **From scientific choices to checkable conclusions.** The agent team (top) plans the test and cohort, calls instruments, specifies the analysis, and interprets the results; a critic reviews the two middle phases. The backend (bottom) is fixed code: it runs the instruments, computes statistics, and derives the evidence label. Its measurement layer rejects implausible readings, compares instruments, and estimates the measurement error that changes significance or effect direction (counterfactual). Interpretation sees the statistics and this assessment, not the label. Colored arrows carry each check's input to the risk level (Eq. 3); no check uses reference answers.

Multi-agent workflow. Role-specialized agents coordinate four phases. Planning and interpretation use agent meetings; a critic reviews imaging and statistics in bounded correction rounds.

1. **Planning.** Agents specify the population, quantities, test family, and covariates, or determine that the hypothesis is untestable. The instrument is fixed before the run; agents choose what to measure, not which instrument measures it.
2. **Image analysis.** An agent requests measurements for the required subjects, structures, and observations. Imaging tools execute the requests: segmentation models for MRI, and a retinal classifier and a vessel-segmentation model for fundus images. The framework records masks, readings, and plausibility assessments.
3. **Statistical analysis.** The framework supplies a base table. The agent assembles analysis variables and a typed specification; framework code applies cohort rules, checks inputs, and computes statistics.
4. **Interpretation.** Agents use the statistics and measurement records to write a verdict and explanation, with the evidence label withheld.

Clinical metadata and human measurements enter the same workflow without segmentation.

Reproducible evidence. An `AnalysisSpec` records statistical choices and a `CohortSpec` defines eligible observations. At a fixed code version, the specification and final analysis table determine the statistics. Seven integrity checks enforce this computation, audit cohort coverage, and scan for fabricated or off-contract data (Appendix C). Agent failures are not retried.

The label rule uses significance, direction, effect size, and interval exclusion (Hoenig & Heisey, 2001), with planned-power criteria where an appropriate interval is unavailable (Appendix B).

Measurement information through interpretation. Where secondary readings exist, instruments are compared under the same specification. Gates can invalidate labels for missing expected structures, implausibly small volumes, or inconsistent cardiac timepoints (Table 3). The counterfactual shifts a continuous biomarker across two groups or the high and low halves of a predictor.

A measurement record keeps each assessment with the instrument that produced it, without overwriting earlier entries. Interpreting agents receive sensitivity estimates and counts of agreeing instrument analyses alongside statistics. Only gates can change a label; the other assessments inform interpretation and risk. These comparisons, records, and gates form the *measurement layer*, removed together in its ablation. The risk level changes neither labels nor verdicts.

4 Benchmark and evaluation design

Evaluating measurement checks requires hypotheses tested with different instruments and references derived independently of those instruments where possible. Scientific-analysis benchmarks generally score answers on supplied data (Majumder et al., 2025; Wang et al., 2025; Chen et al., 2025; Mitchener et al., 2025). We extend the VERITAS benchmark (Stoffl et al., 2026) with a retinal dataset, an external cardiac dataset, additional hypotheses and measurement tools, and references re-derived through the typed backend. The extension tests how measurements affect a conclusion.

The 279 hypotheses span four datasets and five cohort tables: cardiac MRI from ACDC (Bernard et al., 2018) and M&Ms-2 (Martín-Isla et al., 2023), brain MRI and clinical variables from UCSF-PDGM (Calabrese et al., 2022b), and retinal imaging and clinical data from PAPILA (Kovalyk et al., 2022), per eye and per patient. Six complexity tiers and seven control types cover untestable claims through multivariable survival (Table 4). ACDC and UCSF-PDGM come from the VERITAS benchmark; M&Ms-2, absent from the MRI instruments' published training lists, adds sites and vendors; PAPILA adds two human graders. MRI cohorts use up to three segmentation tools; PAPILA provides two human graders, RETFound glaucoma probabilities (Zhou et al., 2023), and VascX vessel measurements (Vargas Quiros et al., 2025). Each cohort's *primary instrument* was fixed before its runs; for UCSF-PDGM and M&Ms-2 it was chosen using reference data (Appendix E.6).

New hypotheses are generated from analysis templates and frozen before computing references. Expert contours or human measurements feed the same backend as live runs. For 22 retinal-model items without expert counterparts, references use the models' outputs on 79 validation eyes, outside classifier training. These items evaluate analysis and interpretation only. An expert clinician checked every hypothesis; Appendix D records construction, exclusions, and inherited items.

Reference decisions and analysis agreement. References depend on the registered analysis and smallest effect of interest (Lakens, 2017; Lakens et al., 2018), anchored in clinical differences, guideline criteria, repeatability, or recorded conventions. We also compare executed specifications with the registered analyses (Table 6): matching a label alone does not establish that the intended hypothesis was tested. This reference-based diagnostic does not enter the risk level.

Systems and outcomes. Two VERITAS-M teams (Sec. 3) are named for their lead model: the *gpt-oss team* uses `gpt-oss:20b`, `qwen3:8b`, and `qwen3-coder:30b` (OpenAI, 2025; Yang et al., 2025; Qwen Team, 2025); the *Qwen3.8 team* uses `qwen3.8:27b` in the agent roles and `gemma4:31b` as critic (Qwen Team, 2026; Gemma Team, 2026).

The *design corpus*, the runs on which we designed the checks, contains 5,210 gpt-oss-team runs, ten repeats per hypothesis in each cohort–instrument pair. The MRI segmentation instruments are SAT, VoxTell, and BiomedParse (Zhao et al., 2025b; Rokuss et al., 2026; Zhao et al., 2025a). Five cohort–instrument pairs use primary instruments. As an instrument stress test, four further pairs

replace the primary instrument with a less accurate one on UCSF-PDGM or M&Ms-2, keeping hypotheses and agents fixed. This *substituted* instrument supplies all of the run's measurements; a *secondary* instrument supplies a comparison. The layer ablation covers all nine pairs.

We chose the checks and threshold $r < 2$ after inspecting the gpt-oss team's design corpus, then fixed them before scoring three transfer systems, all with primary instruments: the single-model (SM) baselines SM-Typed and SM-Code (Sec. 5.1) and the Qwen3.8 team.

We report label accuracy (the computed evidence label) and verdict accuracy (the agents' answer) separately. For selection, a run is *wrong* if either output disagrees with its reference; label-only error tests detection beyond inconsistent interpretation. *Eligible runs* are scorable with a non-INVALID label. Unless stated otherwise, coverage is the share of eligible runs retained; retained error is the share of those runs that are wrong. We also report completion, refusals, and coverage over all attempts. Intervals resample hypotheses, since the ten runs of a hypothesis are dependent.

5 EXPERIMENTS

The experiments test whether VERITAS-M produces better verdicts and whether its checks identify conclusions needing validation. We then ask what measurement checks add beyond consistency and how well the counterfactual estimates sensitivity to known errors. System comparisons use items all compared systems ran; selection and ablations use the pools in Sec. 4.

5.1 ACCURACY OF AUTONOMOUS HYPOTHESIS TESTING

VERITAS-M's advantage over single-model (SM) agents is larger for verdicts than for computed evidence labels (Table 1). The baselines use `qwen3.8:27b` (Qwen Team, 2026): SM-Code writes statistical code, SM-Typed fills the same typed specification as VERITAS-M, and SM-Code and SM-Typed+2 use two correction rounds. With this primary model and a 31B critic, VERITAS-M improves verdict accuracy over SM-Typed+2 by 11.7–25.7 points across five cohort tables; paired intervals exclude zero on four. Label accuracy differs by -2.6 to $+4.7$ points. The comparison also changes the critic, information flow, and computation, leaving their individual contributions unresolved. The gpt-oss team, used to develop the selection rule, has lower label accuracy than SM-Typed+2 on all five tables but higher verdict accuracy on four.

Against VERITAS with the same gpt-oss team, verdict accuracy rises from 81.6 to 93.0% on ACDC items kept from its benchmark and 48.1 to 71.9% on UCSF-PDGM. Counting missing verdicts as wrong gives 72.7 versus 79.3% and 32.3 versus 59.4% (Appendix E.2).

Typed specifications improve label accuracy over free-form code on all five tables (SM-Typed+2 versus SM-Code), but agents can still misinterpret that evidence. Holding interpretation inputs fixed, replacing the Qwen3.8 team's meeting with one model call changes verdict accuracy by at most 0.7 points and preserves $2{,}466/2{,}526$ verdicts (Appendix E.3). The meeting alone therefore does not explain the system's advantage; the contributions of its prompts, earlier phases, and measurement information remain unresolved.

5.2 SELECTING CONCLUSIONS WITH LOWER OBSERVED ERROR

Higher verdict accuracy still leaves conclusions that require validation. To test whether the checks identify them, we retain runs below a risk threshold and measure their error. With primary instruments (2,308 eligible runs), $r < 2$ lowers combined error from 15.3% to 1.5% [0.6, 2.5], retaining 61% and excluding 94% of wrong answers. Label error falls from 6.1% to 1.0%. No error is observed among 651 level-0 runs. Over all attempts, retention is 50%.

The risk level outperforms stated confidence (Fig. 3a): area under the risk–coverage curve (AURC; lower is better) is 0.031 [0.021, 0.043] versus 0.110 [0.080, 0.143]. It closes 87% of the gap between uninformative (random) and oracle ordering (correct conclusions first). Aggregating repeats retains 159 of 262 hypotheses at 1.3% error (two wrong), versus 14.9% before selection (Appendix E.4).

Transfer across systems. After fixing the rule, we scored three further systems on the same benchmark. At $r < 2$, SM-Typed retains 64% at 3.6% error, SM-Code 71% at 6.1%, and the Qwen3.8

Table 1: **Evidence-label / verdict accuracy (%).** Label: computed from the statistics by a fixed rule; verdict: the agent's conclusion. Primary instruments; n hypotheses, ten repeats each. Top: items kept from the VERITAS benchmark; VERITAS rarely reports the interval the label rule needs (—). Middle: items every system ran (all: Table 8); bold: best label and verdict per column. Bottom: verdict gain over SM-Typed+2 with paired 95% bootstrap intervals. Denominators: Table 7.

System	ACDC	UCSF-PDGM	M&Ms-2	PAPILA eyes	PAPILA patients
VERITAS benchmark	$n = 30$	$n = 31$			
VERITAS (gpt-oss)	— / 81.6	— / 48.1	*not in the VERITAS benchmark*		
VERITAS-M (gpt-oss)	93.0 / 93.0	76.2 / 71.9			
Extended benchmark	$n = 64$	$n = 62$	$n = 57$	$n = 18$	$n = 67$
SM-Code	89.2 / 83.1	38.3 / 68.2	81.4 / 68.2	56.7 / 67.8	73.6 / 61.6
SM-Typed	98.4 / 79.8	51.0 / 54.5	79.5 / 70.8	84.0 / 73.3	88.1 / 66.7
SM-Typed+2	**99.8** / 81.5	83.7 / 66.9	84.2 / 73.2	91.3 / 80.0	**91.5** / 70.3
VERITAS-M (gpt-oss)	92.3 / 95.3	76.8 / 69.9	79.1 / 83.7	78.0 / 79.4	86.6 / 84.6
VERITAS-M (Qwen3.8)	97.2 / **99.1**	**85.0 / 83.8**	**84.7 / 84.9**	**96.0 / 98.9**	90.8 / **96.0**
Verdict gain (Qwen3.8)	+17.5	+16.8	+11.7	+18.9	+25.7
95% interval	[9.2, 26.6]	[8.3, 26.6]	[−0.1, 24.1]	[2.8, 37.7]	[15.5, 36.2]

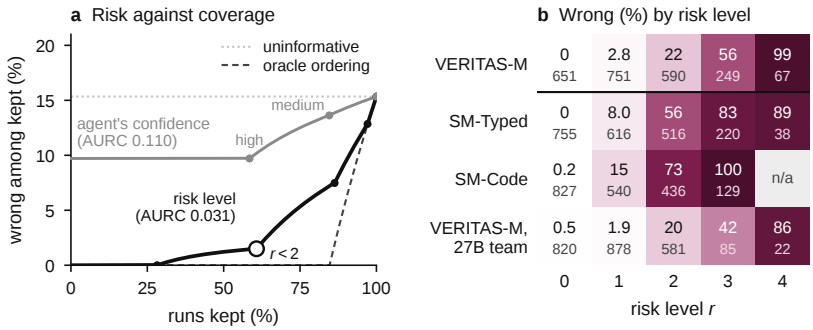

Figure 3: **Selection by risk level and transfer across systems.** Each cohort with its primary instrument; r counts the four checks that flag a run. (a) VERITAS-M, gpt-oss team: error among runs kept in order of increasing risk level (black) or decreasing agent confidence (grey; the 58% rated high tie). AURC: area under the curve, lower is better. Open circle: $r < 2$ keeps 61% at 1.5% error. (b) Percent wrong at each level, run counts below; rows below the line are systems the rule was not tuned on. SM-Code (one model, free code) has no measurement record, so $r = 4$ is n/a.

team 71% at 1.2%, against 25.9, 27.4, and 8.1% before selection. Level-0 errors are $0/755$, $2/827$, and $4/820$ (Fig. 3b); the last four repeat one PAPILA hypothesis with reversed group coding. A threshold chosen offline to target 5% error on the design corpus keeps 1.2 to 9.6% on other systems and instruments, so a useful ranking does not fix a threshold's error rate (Appendix E.5).

5.3 Instrument effects and the contribution of each check

The selection results combine measurement and consistency checks. To isolate what measurement checks add when readings are less accurate, we substitute instruments with hypotheses and agents fixed. M&Ms-2 label accuracy is 78.5% with BiomedParse, 32.5% with SAT, and 14.2% with VoxTell, including gate refusals; the Qwen3.8 team reaches 41.4% with SAT. Table 2 evaluates detection of wrong evidence labels.

The necrosis example shows how such errors distort a scientific contrast. Measuring 495 subjects with each instrument outside any agent run gives a grade-4 versus grade-2 rank-biserial effect of 0.199 with SAT, against 0.906 with expert contours (Table 9). SAT reports necrosis in lower-grade tumors whose expert contours record none, reducing the group difference. Repeating an analysis on those readings preserves the distorted contrast.

With primary instruments, $213/252$ label-conflict flags identify a correct label with a contradicting verdict. The label-conflict and repeat-disagreement checks therefore rank combined error almost as

Table 2: **Detecting wrong evidence labels.** gpt-oss team. Label-error rate (%) in flagged and unflagged runs; recall: share of wrong labels flagged (%). Primary instruments: $141/2{,}308$ labels wrong (6.1%). Substituted instruments (replacing the primary one on UCSF-PDGM or M&Ms-2): $391/1{,}557$ (25.1%). At least one check flags every wrong label. Drop-one results: Appendix E.7.

	Primary instruments			Substituted instruments		
Check	Flagged	Unflagged	Recall	Flagged	Unflagged	Recall
Instrument	13.7	3.0	65	43.2	7.2	86
Label conflict	11.9	5.4	21	31.4	23.8	22
Thin margin	10.1	1.5	89	35.3	6.8	90
Repeat disagreement	13.0	2.5	73	31.8	19.4	59
Any check ($r \geq 1$)	8.5	0.0	100	29.0	0.0	100
Two or more ($r \geq 2$)	14.0	1.0	90	37.9	5.1	92

well as all four (AURC 0.039 versus 0.031). They are weaker on label error (0.031 versus 0.016), leaving 34 wrong labels from 10 hypotheses among $1{,}473$ unflagged runs. Adding instrument and thin-margin checks leaves none among 651; this also reduces coverage.

With substituted instruments, $391/570$ wrong runs have a wrong label. The instrument check detects 86% of wrong labels; flagged runs have 43% label error versus 7% for unflagged runs (Table 2). Removing this check raises label-error AURC from 0.104 to 0.148 and returns 21 wrong labels from three hypotheses to level 0. Consistency checks alone leave 136 wrong labels from 31 hypotheses unflagged. At $r < 2$, all four checks retain 39% with 9.6% combined error, down from 36.6%. Measurement checks therefore add information where consistent interpretations can still be wrong.

The instrument check's two active conditions have different strengths. With substituted instruments, label disagreement has higher label-error precision than counterfactual sensitivity (70% versus 40%). For the Qwen3.8 team, the counterfactual detects 88% of wrong labels; removing it raises retained label error from 0.4 to 1.2% (Table 10).

5.4 Effect of the measurement layer

Whereas removing a check changes selection on saved runs, removing the measurement layer also changes execution. We rerun all nine pairs without gates, instrument comparisons, or measurement information for interpretation. This assesses their combined effect across different runs and eligible pools. At $r < 2$, retained error rises from 4% to 11% and coverage from 52% to 68% of each condition's eligible runs (Table 11). Level 0 contains $20/1{,}198$ wrong runs versus $0/859$ with the layer; all 20 concern three hypotheses with substituted instruments. At matched coverage, estimates favor the layer on UCSF-PDGM with SAT (11% versus 15%) and M&Ms-2 with VoxTell (22% versus 35%) and disfavor it by at most 2.2 points on three pairs, all within wide intervals (Table 12).

With primary instruments, enabling the layer changes label accuracy by -5.8 to $+0.2$ percentage points and verdict accuracy by -2.3 to $+1.2$ points. Larger label-accuracy losses with substituted M&Ms-2 instruments reflect gate refusals (Table 11). Gates reject implausible readings that can still yield the reference answer, so accuracy and selective error capture different outcomes.

5.5 Testing the counterfactual with known errors

We next test whether the counterfactual accurately estimates decision-changing error. We inject differential shifts into expert readings for $141/271$ typed hypotheses with characterized continuous biomarkers. A fivefold finer grid locates full-label changes; 39 items show none within the searched range. Among the other 102, the full label changes at a median 90% of the reported counterfactual (ratio 0.90 [0.86, 0.95]; 5th–95th percentiles 0.20–2.40). The counterfactual therefore overstates tolerated error at the median. Grid spacing explains the systematic shortfall; differences between significance and full-label decisions account for further variation across items (Fig. 6).

With matched assumed and injected scales near one error unit, the check detects 84–87% of label changes at 2–4% false alarms. Assuming a scale half the injected bias detects about two thirds of changes; a scale 1.5 times the bias detects 99% but flags about one fifth of unchanged labels

(Appendix E.9). Detection therefore depends on the assumed scale; the counterfactual does not certify stability. Appendix E.10 tests effect-size thresholds with verdicts held fixed.

6 FAILURE MODES AND LIMITATIONS

Consistent analysis errors remain difficult to detect. Wrong variables, populations, or group coding can produce reproducible statistics and unflagged verdicts across repeats. Specification agreement identifies some mismatches offline but requires the registered analysis. A correct verdict can also accompany a wrong label because several labels map to the same verdict (Appendix F).

Instrument agreement cannot reveal shared error. SAT and VoxTell errors correlate at 0.70–0.87 across four UCSF-PDGM structures. Their disagreement therefore understates dispersion relative to expert contours (Appendix H.1). The counterfactual also covers only specified differential shifts in continuous biomarkers.

Missing checks contribute zero, so unavailable assessments can lower the risk level. The repeat check requires three executions per hypothesis; without it, 11 wrong runs from 10 hypotheses remain at level 0 with primary instruments. Across all attempts, $r < 2$ retains 39% over nine pairs and 50% with primary instruments. Thresholds need validation under changed systems or instruments.

The four clinical datasets partly overlap instrument training data. M&Ms-2 tests external sites, but UCSF-PDGM's primary instrument was selected using its own expert contours. References were withheld at run time but used in development, selection, and evaluation, and carry annotation error themselves (Appendix G). An expert clinician reviewed every hypothesis; transfer to other scientific domains remains untested.

7 DISCUSSION

We asked how an autonomous science agent can identify unreliable conclusions without reference answers. Our answer is to treat its tools as instruments: compare them, test whether assumed measurement error could change the decision, and add checks for contradictions, thin margins, and unstable repeats. Conclusions flagged by at most one check were far less often wrong on every system we evaluated.

Measurement checks reveal failures that consistent reasoning can preserve. In the instrument stress test, inaccurate readings often yield a wrong evidence label that the agent interprets consistently. Comparing instruments and testing measurement sensitivity expose some of these errors, complementing checks for contradictions and disagreement across runs.

Measurement dispersion alone does not determine which conclusions are wrong. Expert-derived error scales on UCSF-PDGM flag more runs without improving error ranking (Appendix H.2). Reporting sensitivity makes explicit how measurement error affects the particular conclusion.

System accuracy and error selection also measure different outcomes. VERITAS-M improves verdict accuracy, while its measurement layer reduces error among selected conclusions. The layer can reject implausible readings that happen to produce a reference-matching answer, so accuracy, refusal, and retained error must be judged together. The small benefit of the interpretation meeting further motivates studying which information agents receive, alongside how they collaborate.

An autonomous workflow can make reliability reporting routine for each conclusion: the instruments behind its measurements, sensitivity in units of assumed measurement error, stability across repeats, and a risk level. Many-analyst studies demonstrate the value of examining how analysis choices affect conclusions (Silberzahn et al., 2018; Botvinik-Nezer et al., 2020), although repeated agent runs are not independent analysts. With primary instruments, the 39% of eligible runs at $r \geq 2$ contain 94% of observed wrong answers, so review can focus on them. In clinical research, the flags can guide validation: examine measurements after instrument disagreement, check interpretation after a label conflict, and inspect analysis choices after repeat disagreement. A thin margin calls for assessing whether the available precision resolves the effect of interest. For AI research, they make the ability to recognize unreliable evidence an explicit evaluation target alongside answer accuracy.

Reproducibility statement

We will release the code, the benchmark with its frozen references, the agent prompts, and the run artifacts upon publication. Saved specifications and analysis tables allow statistics to be regenerated; frozen references can be recomputed from expert-derived data and, for the retinal-model items, from the saved model outputs. Appendix B defines the label rule and checks, Appendices C–D describe the implementation and benchmark, and Appendix E reports the full experimental results.

Ethics statement

The four datasets are public, de-identified research cohorts used under their licenses; no new patient data were collected. Images and masks remain within the compute environment. All experiments use locally deployable models, including a 31B critic in the Qwen3.8 team. Reproducible analyses can still support incorrect clinical conclusions. These findings support research assistance with expert review, not autonomous clinical use.

AI use statement

We used generative AI tools for developing the formal framework, experiment execution, code development, data analysis and interpretation, literature search, figure drafting, and manuscript drafting and editing under author direction. Reported results were computed by the described experimental pipeline. The authors checked the AI-assisted contributions against the code, data, and result files, and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Rahim Abo Kasem, Lydia A. Leavitt, Gina Genova, Brian N. Snyder, Isaac J. Abecassis, Adrianna H. Masters, Guneet Sarai, Brian J. Williams, Jonathan Beall, and Akshitkumar M. Mistry. Biostatistical benchmarking of neuro-oncology trials. *Neuro-Oncology Advances*, 8(1):vdaf257, 2026. doi: 10.1093/noajnl/vdaf257.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000.

Laura Battaglia, Timothy Christensen, Stephen Hansen, and Szymon Sacher. Inference for regression with variables generated by AI or machine learning. *arXiv preprint arXiv:2402.15585*, 2024.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del Arco, Johannes B. Gruber, and Dirk Hovy. Large language model hacking: Quantifying the hidden risks of using LLMs for text annotation. *arXiv preprint arXiv:2509.08825*, 2025.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018. doi: 10.1109/TMI.2018.2837502.

Martin Bertran, Riccardo Fogliato, and Zhiwei Steven Wu. Many AI analysts, one dataset: Navigating the agentic data science multiverse. *Proceedings of the National Academy of Sciences*, 123 (29):e2606495123, 2026. doi: 10.1073/pnas.2606495123.

J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986. doi: 10.1016/S0140-6736(86)90837-8.

Thierry Blankenstein, Jialin Yu, Zixuan Li, Vassilis Plachouras, Sunando Sengupta, Philip Torr, Yarin Gal, Alasdair Paren, and Adel Bibi. BiasBusters: Uncovering and mitigating tool selection bias in large language models. In *International Conference on Learning Representations (ICLR)*, 2026.

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.

Rotem Botvinik-Nezer, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A. Mumford, R. Alison Adcock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810): 84–88, 2020. doi: 10.1038/s41586-020-2314-9.

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.

Matthias Bruhns, Jan T. Schleicher, Maximilian Wirth, Marcello Zago, Sepideh Babaei, and Manfred Claassen. Effects of segmentation errors on downstream-analysis in highly-multiplexed tissue imaging. *PLOS Computational Biology*, 21(9):e1013350, 2025. doi: 10.1371/journal.pcbi.1013350.

Evan Calabrese, Javier Villanueva-Meyer, Jeffrey Rudie, Andreas Rauschecker, Ujjwal Baid, Spyridon Bakas, Soonmee Cha, John Mongan, and Christopher Hess. The University of California San Francisco preoperative diffuse glioma MRI (UCSF-PDGM). The Cancer Imaging Archive, version 5, 2022a. Dataset.

Evan Calabrese, Javier E. Villanueva-Meyer, Jeffrey D. Rudie, Andreas M. Rauschecker, Ujjwal Baid, Spyridon Bakas, Soonmee Cha, John T. Mongan, and Christopher P. Hess. The University of California San Francisco preoperative diffuse glioma MRI dataset. *Radiology: Artificial Intelligence*, 4(6):e220058, 2022b. doi: 10.1148/ryai.220058.

Víctor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martín-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021. doi: 10.1109/TMI.2021.3090082.

Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Number 105 in Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, Boca Raton, FL, 2 edition, 2006. ISBN 978-1-58488-633-4. doi: 10.1201/9781420010138.

Rickey E. Carter, Paul M. McKie, and Curtis B. Storlie. The Fragility Index: a P-value in sheep's clothing? *European Heart Journal*, 38(5):346–348, 2017. doi: 10.1093/eurheartj/ehw495.

Balwantray C. Chauhan, David F. Garway-Heath, Francisco J. Goñi, Luca Rossetti, Boel Bengtsson, Ananth C. Viswanathan, and Anders Heijl. Practical recommendations for measuring rates of visual field change in glaucoma. *British Journal of Ophthalmology*, 92(4):569–573, 2008. doi: 10.1136/bjo.2007.135012.

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. In *International Conference on Learning Representations (ICLR)*, 2025.

Matt Y Cheung, Tucker J Netherton, Laurence E Court, Ashok Veeraraghavan, and Guha Balakrishnan. Bias-aware conformal prediction for metric-based imaging pipelines. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2026a. doi: 10.1109/ISBI61048.2026.11515847.

Matt Y Cheung, Ashok Veeraraghavan, and Guha Balakrishnan. ConRad: Efficient conformal prediction for radiomics. *arXiv preprint arXiv:2607.08084*, 2026b.

Zach Eaton-Rosen, Felix Bragman, Sotirios Bisdas, Sébastien Ourselin, and M Jorge Cardoso. Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 11070 of *Lecture Notes in Computer Science*, pp. 691–699, 2018. doi: 10.1007/978-3-030-00928-1_78.

Steven Young Eulig. Position: Correct answer, wrong mechanism – when AI scientists defend general claims their own data contradicts. *arXiv preprint arXiv:2606.23175*, 2026.

Dyke Ferber, Lars Hilgers, Christiane Höper, Benedict Kinny-Köster, Jan-Niklas Eckardt, Katharina Egger-Heidrich, Marius Bill, Martin M. K. Schneider, Jan Clusmann, Lejla Kadric, Marcel Oehme, Maximilian Mayrhofer-Schmid, Alexander Oeser, Georg Wölflein, Isabella C. Wiest, Jan Moritz Middeke, A. John Iafrate, Daniel Truhn, Dirk Jäger, and Jakob Nikolas Kather. Towards autonomous medical artificial intelligence agents. *Nature*, 655(8125):1282–1291, 2026. doi: 10.1038/s41586-026-10675-5.

Pavel Filip, Petr Bednarik, Lynn E. Eberly, Amir Moheet, Alena Svatkova, Heidi Grohn, Anjali F. Kumar, Elizabeth R. Seaquist, and Silvia Mangia. Different FreeSurfer versions might generate different statistical outcomes in case–control comparison studies. *Neuroradiology*, 64(4):765–773, 2022. doi: 10.1007/s00234-021-02862-0.

Shanghua Gao, Richard Zhu, Pengwei Sui, Zhenglun Kong, Sufian Aldogom, Yepeng Huang, Ayush Noori, Reza Shamji, Krishna Parvataneni, Theodoros Tsiligkaridis, and Marinka Zitnik. ToolUniverse: An open platform for democratizing AI scientists. *arXiv preprint arXiv:2509.23426*, 2025. doi: 10.48550/arXiv.2509.23426.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Lena Giebeler, Deepa Krishnaswamy, David Clunie, Jakob Wasserthal, Lalith Kumar Shiyam Sundar, Andres Diaz-Pinto, Klaus H. Maier-Hein, Murong Xu, Bjoern Menze, Steve Pieper, Ron Kikinis, and Andrey Fedorov. In search of truth: Evaluating concordance of AI-based anatomy segmentation models. *Journal of Medical Imaging*, 13(6):062204, 2026. doi: 10.1117/1.jmi.13.6.062204.

Nayoon Gim, In Gim, Yu Jiang, Yuka Kihara, Marian Blazes, Yue Wu, Cecilia S Lee, and Aaron Y Lee. An LLM-assisted framework for accelerated and verifiable clinical hypothesis testing from electronic health records. *medRxiv*, 2026. doi: 10.64898/2026.02.10.26346008.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, 2026. doi: 10.1038/s41586-026-10644-y.

Frank Grothues, Gillian C. Smith, James C. C. Moon, Nicholas G. Bellenger, Peter Collins, Helmut U. Klein, and Dudley J. Pennell. Comparison of interstudy reproducibility of cardiovascular magnetic resonance with two-dimensional echocardiography in normal subjects and in patients with heart failure or left ventricular hypertrophy. *American Journal of Cardiology*, 90(1):29–34, 2002. doi: 10.1016/S0002-9149(02)02381-0.

Frank Grothues, James C. Moon, Nicholas G. Bellenger, Gillian S. Smith, Helmut U. Klein, and Dudley J. Pennell. Interstudy reproducibility of right ventricular volumes, function, and mass with cardiovascular magnetic resonance. *American Heart Journal*, 147(2):218–223, 2004. doi: 10.1016/j.ahj.2003.10.005.

Frank E. Grubbs. On estimating precision of measuring instruments and product variability. *Journal of the American Statistical Association*, 43(242):243–264, 1948. doi: 10.1080/01621459.1948.10483261.

Bhaskar Gurram. Auditing automated evaluation, error propagation, and runtime mitigation in tool-using language agents. *arXiv preprint arXiv:2604.16706*, 2026.

Monika E. Hegi, Annie-Claire Diserens, Thierry Gorlia, Marie-France Hamou, Nicolas de Tribolet, Michael Weller, Johan M. Kros, Johannes A. Hainfellner, Warren Mason, Luigi Mariani, Jacoline E. C. Bromberg, Peter Hau, René O. Mirimanoff, J. Gregory Cairncross, Robert C. Janzer, and Roger Stupp. MGMT gene silencing and benefit from temozolomide in glioblastoma. *New England Journal of Medicine*, 352(10):997–1003, 2005. doi: 10.1056/NEJMoa043331.

John M Hoenig and Dennis M Heisey. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1):19–24, 2001. doi: 10.1198/000313001300339897.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candès, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25372–25437, 2025.

S. L. Hui and S. D. Walter. Estimating the error rates of diagnostic tests. *Biometrics*, 36(1):167–171, 1980. doi: 10.2307/2530508.

Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. Autonomous LLM-driven research—from data to human-verifiable research papers. *NEJM AI*, 2(1):AIoa2400555, 2025. doi: 10.1056/AIoa2400555.

Joint Committee for Guides in Metrology. Evaluation of measurement data — guide to the expression of uncertainty in measurement. Guide JCGM 100:2008, Joint Committee for Guides in Metrology (JCGM), published by BIPM, Sèvres, France, 2008.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=Zy4uFzMviZ.

Nadine Kawel-Boehm, Scott J. Hetzel, Bharath Ambale-Venkatesh, Gabriella Captur, Christopher J. Francois, Michael Jerosch-Herold, Michael Salerno, Shawn D. Teague, Emanuela Valsangiacomo-Buechel, Rob J. van der Geest, and David A. Bluemke. Reference ranges ("normal values") for cardiovascular magnetic resonance (CMR) in adults and children: 2020 update. *Journal of Cardiovascular Magnetic Resonance*, 22(1):87, 2020. doi: 10.1186/s12968-020-00683-3.

Simon Klau, Sabine Hoffmann, Chirag J. Patel, John P. A. Ioannidis, and Anne-Laure Boulesteix. Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework. *International Journal of Epidemiology*, 50 (1):266–278, 2021. doi: 10.1093/ije/dyaa164.

Joshua R. Klein and Aaron Roodman. Blind analysis in nuclear and particle physics. *Annual Review of Nuclear and Particle Science*, 55(1):141–163, 2005. doi: 10.1146/annurev.nucl.55.090704.151521.

Florian Kofler, Ivan Ezhov, Lucas Fidon, Carolin M. Pirkl, Johannes C. Paetzold, Egon Burian, Sarthak Pati, Malek El Husseini, Fernando Navarro, Suprosanna Shit, Jan Kirschke, Spyridon Bakas, Claus Zimmer, Benedikt Wiestler, and Bjoern H. Menze. Robust, primitive, and unsupervised quality estimation for segmentation ensembles. *Frontiers in Neuroscience*, 15:752780, 2021. doi: 10.3389/fnins.2021.752780.

Aachal Kotecha, Edward White, Patricio G. Schlottmann, and David F. Garway-Heath. Intraocular pressure measurement precision with the Goldmann applanation, dynamic contour, and ocular response analyzer tonometers. *Ophthalmology*, 117(4):730–737, 2010. doi: 10.1016/j.ophtha.2009.09.020.

Oleksandr Kovalyk, Juan Morales-Sánchez, Rafael Verdú-Monedero, Inmaculada Sellés-Navarro, Ana Palazón-Cabanes, and José-Luis Sancho-Gómez. PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Scientific Data*, 9:291, 2022. doi: 10.1038/s41597-022-01388-1.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

Daniël Lakens, Anne M Scheel, and Peder M Isager. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269, 2018. doi: 10.1177/2515245918770963.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, 3rd edition, 2005. ISBN 978-0-387-98864-1. doi: 10.1007/0-387-27605-X.

Hongmin Li. The calibration turn in AI-assisted research: A conceptual and methodological framework for evidence-licensed claims. *arXiv preprint arXiv:2606.31273*, 2026.

Jiameng Li, Teodora Popordanoska, Aleksei Tiulpin, Sebastian G Gruber, Frederik Maes, and Matthew B Blaschko. CARE: Confidence-aware ratio estimation for medical biomarkers. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 172–197. PMLR, 2026.

Junqi Liu, Selena Song, Yuhan Wang, Jiawei Mao, Hardy Chen, Xiaoke Huang, Tianhao Qi, Pengfei Guo, Yucheng Tang, Yufan He, Can Zhao, Andriy Myronenko, Dong Yang, Daguang Xu, and Yuyin Zhou. AutoMedBench: Towards medical AutoResearch with agentic AI models. *arXiv preprint arXiv:2606.01961*, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651(8107):914–919, 2026. doi: 10.1038/s41586-026-10265-5.

Ziming Luo, Atoosa Kasirzadeh, and Nihar B. Shah. The more you automate, the less you see: Hidden pitfalls of AI scientist systems. In *NeurIPS 2025 AI4Science Workshop*, 2025.

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. In *International Conference on Learning Representations (ICLR)*, 2025.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York, 2003. ISBN 978-0-387-00454-9. doi: 10.1007/b97478.

Carlos Martín-Isla, Víctor M. Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, others, and Karim Lekadir. Deep learning segmentation of the right ventricle in cardiac MRI: The M&Ms challenge. *IEEE Journal of Biomedical and Health Informatics*, 27 (7):3302–3313, 2023. doi: 10.1109/JBHI.2023.3267857.

Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878, 2020. doi: 10.1109/TMI.2020.3006437.

Jiacheng Miao, Jonathan K. Pritchard, and James Zou. The agentic garden of forking paths. *arXiv preprint arXiv:2607.01507*, 2026.

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.

Changdae Oh, Seongheon Park, To Eun Kim, Jiatong Li, Wendi Li, Samuel Yeh, Sean Du, Hamed Hassani, Paul Bogdan, Dawn Song, and Sharon Li. Uncertainty quantification in LLM agents: Foundations, emerging challenges, and opportunities. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16219–16250, 2026. doi: 10.18653/v1/2026.acl-long.738.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Nathan Painchaud, Youssef Skandarani, Thierry Judge, Olivier Bernard, Alain Lalande, and Pierre-Marc Jodoin. Cardiac segmentation with strong anatomical guarantees. *IEEE Transactions on Medical Imaging*, 39(11):3703–3713, 2020. doi: 10.1109/TMI.2020.3003240.

Steffen E. Petersen, Nay Aung, Mihir M. Sanghvi, Filip Zemrak, Kenneth Fung, Jose Miguel Paiva, Jane M. Francis, Mohammed Y. Khanji, Elena Lukaschuk, Aaron M. Lee, Valentina Carapella, Young Jin Kim, Paul Leeson, Stefan K. Piechnik, and Stefan Neubauer. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. *Journal of Cardiovascular Magnetic Resonance*, 19(1): 18, 2017. doi: 10.1186/s12968-017-0327-9.

Qwen Team. Qwen3-Coder-30B-A3B-Instruct. `https://huggingface.co/Qwen/Qwen3-Coder-30B-A3B-Instruct`, 2025. Model card, released 31 July 2025. Accessed 2026-09-25.

Qwen Team. Qwen3.8-27B. `https://huggingface.co/Qwen/Qwen3.8-27B`, 2026. Model card. Accessed 2026-09-24.

Harsh Raj, Niranjan Orkat, Suvrorup Mukherjee, Aritra Guha, Cheryl Flynn, and Subhabrata Majumdar. Consistency as a testable property: Statistical methods to evaluate AI agent reliability. *arXiv preprint arXiv:2605.10516*, 2026.

Zachary T. Rewolinski, Austin V. Zane, Hao Huang, Chandan Singh, Chenglong Wang, Jianfeng Gao, and Bin Yu. Sanity checks for agentic data science. *arXiv preprint arXiv:2604.11003*, 2026.

Maximilian Rokuss, Moritz Langenberg, Yannick Kirchhoff, Fabian Isensee, Benjamin Hamm, Constantin Ulrich, Sebastian Regnery, Lukas Bauer, Efthimios Katsigiannopulos, Tobias Norajitra, and Klaus Maier-Hein. VoxTell: Free-text promptable universal 3D medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37538–37557, 2026.

Paul R. Rosenbaum. *Observational Studies*. Springer Series in Statistics. Springer, New York, 2nd edition, 2002. ISBN 978-1-4419-3191-7. doi: 10.1007/978-1-4757-3692-2.

Karen Sargsyan. Structural enforcement of statistical rigor in AI-driven discovery: A functional architecture. *arXiv preprint arXiv:2511.06701*, 2025. doi: 10.48550/arXiv.2511.06701.

Kaiwen Shi, Zheyuan Zhang, Han Bao, Colby Nelson, and Yanfang Ye. Confidence laundering in agent systems: Why uncertainty needs a latent carrier. *arXiv preprint arXiv:2606.20662*, 2026.

Raphael Silberzahn, Eric L. Uhlmann, Daniel P. Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štěpán Bahník, Feng Bai, Colin Bannard, Evelina Bonnier, et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 2018. doi: 10.1177/2515245917747646.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637.

Lucas Stoffl, Benedikt Wiestler, and Johannes C Paetzold. Veritas: A multi-agent co-scientist for verifiable image-derived hypothesis testing. In *European Conference on Computer Vision*, pp. 502–520. Springer, 2026. doi: 10.1007/978-3-032-37362-5_27.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, 2025. doi: 10.1038/s41586-025-09442-9.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023. doi: 10.52202/075280-3275.

Michelle Vaccaro. Position: Preregister experiments with AI agents. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026.

Vanya V. Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, and Ben Glocker. Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth. *IEEE Transactions on Medical Imaging*, 36(8):1597–1606, 2017. doi: 10.1109/TMI.2017.2665165.

Tyler J. VanderWeele and Peng Ding. Sensitivity analysis in observational research: Introducing the E-value. *Annals of Internal Medicine*, 167(4):268–274, 2017. doi: 10.7326/M16-2607.

Tyler J VanderWeele and Yige Li. Simple sensitivity analysis for differential measurement error. *American Journal of Epidemiology*, 188(10):1823–1829, 2019. doi: 10.1093/aje/kwz133.

Jose Vargas Quiros, Bart Liefers, Karin A. van Garderen, Jeroen P. Vermeulen, and Caroline Klaver. VascX models: Deep ensembles for retinal vascular analysis from color fundus images. *Translational Vision Science & Technology*, 14(7):19, 2025. doi: 10.1167/tvst.14.7.19.

Rohit Varma, William C. Steinmann, and Ingrid U. Scott. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology*, 99(2):215–221, 1992. doi: 10.1016/S0161-6420(92)31990-6.

Michael Walsh, Sadeesh K. Srinathan, Daniel F. McAuley, Marko Mrkobrada, Oren Levine, Christine Ribic, Amber O. Molnar, Neil D. Dattani, Andrew Burke, Gordon Guyatt, Lehana Thabane, Stephen D. Walter, Janice Pogue, and P. J. Devereaux. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. *Journal of Clinical Epidemiology*, 67(6):622–628, 2014. doi: 10.1016/j.jclinepi.2013.10.019.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019. doi: 10.1016/j.neucom.2019.01.103.

Siruo Wang, Tyler H. McCormick, and Jeffrey T. Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117 (48):30266–30275, 2020. doi: 10.1073/pnas.2001238117.

Zhongyuan Wang and Pratyusha Vemuri. When the tool decides: LLM agents defer blindly to graph neural network tools, and stronger backbones defer more. *arXiv preprint arXiv:2606.14476*, 2026.

Zifeng Wang, Benjamin Danek, and Jimeng Sun. BioDSA-1K: Benchmarking data science agents for biomedical research. *arXiv preprint arXiv:2505.16100*, 2025. doi: 10.48550/arXiv.2505.16100.

Patrick Y. Wen, Martin van den Bent, Gilbert Youssef, Timothy F. Cloughesy, Benjamin M. Ellingson, Michael Weller, Evanthia Galanis, Daniel P. Barboriak, John de Groot, Mark R. Gilbert, et al. RANO 2.0: update to the response assessment in neuro-oncology criteria for high- and low-grade gliomas in adults. *Journal of Clinical Oncology*, 41(33):5187–5199, 2023. doi: 10.1200/JCO.23.01059.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Jennifer J. Yland, Amelia K. Wesselink, Timothy L. Lash, and Matthew P. Fox. Misconceptions about the direction of bias from nondifferential misclassification. *American Journal of Epidemiology*, 191(8):1485–1495, 2022. doi: 10.1093/aje/kwac035.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moung-Wen, Brian Piening, Carlo Bifulco, Mu Wei, Hoifung Poon, and Sheng Wang. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods*, 22: 166–176, 2025a. doi: 10.1038/s41592-024-02499-w.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Xiao Zhou, Ya Zhang, Yanfeng Wang, and Weidi Xie. Large-vocabulary segmentation for medical images with text prompts. *npj Digital Medicine*, 8(1):566, 2025b. doi: 10.1038/s41746-025-01964-w.

Lujia Zhong, Yihao Xia, Jianwei Zhang, Shuo Huang, Jiaxin Yue, Mingyang Xia, and Yonggang Shi. NeuroAgent: LLM agents for multimodal neuroimaging analysis and research. *arXiv preprint arXiv:2605.06584*, 2026.

Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, Mateo G. Lozano, Peter Woodward-Court, Yuka Kihara, UK Biobank Eye & Vision Consortium, Andre Altmann, Aaron Y. Lee, Eric J. Topol, Alastair K. Denniston, Daniel C. Alexander, and Pearse A. Keane. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023. doi: 10.1038/s41586-023-06555-x.

Ivan Zhovannik, Dennis Bontempi, Alessio Romita, Elisabeth Pfaehler, Sergey Primakov, Andre Dekker, Johan Bussink, Alberto Traverso, and René Monshouwer. Segmentation uncertainty estimation as a sanity check for image biomarker studies. *Cancers*, 14(5):1288, 2022. doi: 10.3390/cancers14051288.

Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024. doi: 10.1073/pnas.2322083121.

Alex Zwanenburg, Martin Vallières, Mahmoud A Abdalah, Hugo JWL Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J Beukinga, Ronald Boellaard, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, 2020. doi: 10.1148/radiol.2020191145.

A Related work

Autonomous science and reproducible workflows. Science agents connect language models to experimental tools (Boiko et al., 2023; Bran et al., 2024), coordinate collaborating agents (Gottweis et al., 2026; Swanson et al., 2025), and automate computational research (Lu et al., 2026). Clinical and biomedical systems extend these approaches to diagnostic testing and neuroimaging analysis (Ferber et al., 2026; Zhong et al., 2026). Our setting is hypothesis testing on an existing clinical cohort, before independent measurements or experiments validate the conclusion.

Data-to-paper connects results to executable code and data (Ifargan et al., 2025), while clinical assistants make health-record analyses inspectable (Gim et al., 2026). VERITAS links hypotheses, imaging, statistics, and verdicts (Stoffl et al., 2026); VERITAS-M adds typed computation and measurement checks. These workflows establish how a result was obtained. Assessing measurement accuracy requires further evidence.

Hypothesis validation and evaluation. POPPER validates hypotheses through sequential falsification with Type-I error control under its assumptions (Huang et al., 2025). Our framework asks how errors in the measurements entering a test affect its conclusion. The two address complementary questions: statistical error control does not by itself assess measurement accuracy, and our checks do not provide POPPER's error guarantee.

DiscoveryBench, BioDSA, ScienceAgentBench, and BixBench evaluate scientific analysis against reference tasks (Majumder et al., 2025; Wang et al., 2025; Chen et al., 2025; Mitchener et al., 2025). Medical-AI research benchmarks extend this scope to end-to-end imaging workflows (Liu et al., 2026). BioDSA's distinction between decision accuracy and evidence alignment relates to

our separate labels and verdicts. Controlled audits of AI-scientist systems find benchmark-selection and post-hoc selection biases, undocumented data substitution, and arbitrary metric choices (Luo et al., 2025). Following broader evaluation concerns (Kapoor et al., 2025), we distinguish accuracy, completion, refusals, and error after selection.

Variation across analyses and runs. Agents can reach different conclusions from the same data or alter analyses with their stated beliefs (Bertran et al., 2026; Miao et al., 2026). This connects their evaluation to multiverse analysis, many-analyst studies, and the vibration of effects (Steegen et al., 2016; Silberzahn et al., 2018; Botvinik-Nezer et al., 2020; Klau et al., 2021). Self-consistency, semantic and trajectory uncertainty, and deferral policies also assess outputs (Oh et al., 2026; Raj et al., 2026). Repeating analyses of the same biased measurements can nevertheless preserve the error.

Uncertainty may be lost as outputs pass between agents (Shi et al., 2026). Our measurement records preserve assessments through interpretation. Stated confidence, stability, and fidelity to evidence are distinct (Rewolinski et al., 2026; Eulig, 2026), and reported reasoning need not faithfully explain an answer (Turpin et al., 2023). We therefore compare agent confidence with computed reliability checks.

Tool use and constrained computation. Tool-use methods support selecting and invoking external capabilities (Yao et al., 2023; Gao et al., 2025). Perturbing tool-call parameters assesses one source of sensitivity (Gurram, 2026); our counterfactual perturbs measurements entering the statistical analysis. Agents' tool choices can themselves be biased (Blankenstein et al., 2026), and flexible analysis choices introduce selection effects (Baumann et al., 2025; Vaccaro, 2026), connecting agent workflows to inference after selection (Berk et al., 2013). Fixing instruments before inspecting readings follows the principle of blind analysis (Klein & Roodman, 2005). Architectural constraints can also separate validation data and track statistical error budgets (Sargsyan, 2025). Our typed interface makes the computation reproducible and its inputs inspectable; correct analysis selection remains a separate question.

Measurement without a reference standard. Multi-instrument error estimation (Grubbs, 1948) and diagnostic testing without a gold standard (Hui & Walter, 1980) require assumptions to identify error. Agreement plots characterize differences but cannot identify shared bias (Bland & Altman, 1986). Measurement-error theory, partial identification, and metrology provide the broader context (Carroll et al., 2006; Manski, 2003; Joint Committee for Guides in Metrology, 2008). Our assumed allocation of pairwise error inherits these identification limits.

In imaging, biomarker standardization defines quantities consistently (Zwanenburg et al., 2020), and segmentation errors or software versions can change scientific conclusions (Bruhns et al., 2025; Filip et al., 2022). Concordance between segmentation models provides reference-free agreement information (Giebeler et al., 2026). Human graders can serve as instruments too.

Mask quality and biomarker uncertainty. Segmentation calibration and test-time augmentation estimate voxel uncertainty (Mehrtash et al., 2020; Wang et al., 2019), which can be propagated to derived volumes (Eaton-Rosen et al., 2018). Reverse classification accuracy assesses a mask without a case-specific reference (Valindria et al., 2017). CARE and conformal intervals address uncertainty in derived biomarkers (Li et al., 2026; Cheung et al., 2026b;a). Such estimates could supply measurement information to our workflow; the present counterfactual uses instrument disagreement and assesses the downstream decision.

Inference with predicted variables. Prediction-powered inference and its extensions use labeled references to correct inference with predicted variables (Angelopoulos et al., 2023; Zrnic & Candès, 2024). Post-prediction inference and generated-regressor methods address related questions (Wang et al., 2020; Battaglia et al., 2024). Our runtime checks use instrument disagreement when references are withheld or unavailable. This supplies an assumed sensitivity scale but cannot identify a correction for shared bias. Where reference data exist, correction-based inference can complement these checks.

Sensitivity and selective prediction. Decision-changing perturbations connect our counterfactual to the fragility index (Walsh et al., 2014), causal sensitivity analysis (Rosenbaum, 2002; VanderWeele & Ding, 2017), and differential measurement error (VanderWeele & Li, 2019). Zhovannik et al. (2022) examine contour perturbations and survival conclusions using published inter-observer variation. We scale biomarker perturbations by the compared instruments and recompute through the same backend. Both shift directions matter because measurement error need not attenuate an association (Yland et al., 2022); rank invariance explains why some transformations leave tests unchanged (Lehmann & Romano, 2005). Criticism of significance-based fragility measures (Carter et al., 2017) motivates our comparison between the deployed significance decision and the full evidence label.

Risk–coverage evaluation is established in selective prediction (Geifman & El-Yaniv, 2017; Angelopoulos & Bates, 2023). We use it to assess whether measurement, sampling, and consistency checks select scientific conclusions with lower observed error. The threshold bound is standard; our contribution connects measurement assessments to the computed evidence and tests their value beyond consistency in a clinical co-scientist. Proposals to report reliability alongside scientific outputs (Li, 2026) have precedents in structure prediction (Jumper et al., 2021). Here the reported information concerns measurements, decision sensitivity, and interpretation.

B Formal definitions and proofs

B.1 Definitions

Readings and specifications. A reading records its instrument, version, and source. Where a secondary instrument is available, disagreement with the run's instrument supplies the assumed error scale σ. The specification ϕ names the family, variables, covariates, claimed direction, α, and population. Reference readings are expert or human measurements, except for the retinal-model items defined in Appendix D.

The label rule f. The rule first checks the routes to INVALID: a missing or broken statistic, a degenerate fit, a violated test family or constraint, and the plausibility gates of Sec. 3. Any of these makes the label INVALID. Otherwise f reads four quantities and nothing else:

- *significance*: the run's p-value against the specification's α, which may be stricter than the registered default but never looser; a conjunctive claim uses $\max(p, p_{\text{conj}})$, where p_{conj} is the largest per-predictor p-value;
- *direction*: the observed direction, re-oriented onto the hypothesis's canonical group order, against the direction the hypothesis registers, not the one the agent declares;
- *size*: whether the estimate reaches δ on δ's scale, which is distance from no effect, on the log scale for a ratio;
- *exclusion*: whether the interval on that scale rules out a δ-sized effect, one-sided toward the claimed direction when the hypothesis is directional.

The branches are:

1. A significant effect in the claimed direction is SUPPORTED if it reaches δ, REFUTED if it does not and the interval excludes a δ-sized effect, and UNDERPOWERED otherwise. If the family's effect shares no scale with δ, the size test is skipped and the effect is SUPPORTED.
2. A significant effect in the wrong direction is REFUTED.
3. A null result is REFUTED if the interval excludes a δ-sized effect and UNDERPOWERED otherwise. Families with no interval on δ's scale (categorical associations, the joint test of several regression predictors, and equivalence tests) fall back to the planned power at δ and α: the null is REFUTED if that power is at least 0.8.

So f is piecewise constant in the estimate, the two interval bounds and p, with boundaries at 0 and $\pm\delta$ on δ's scale and at α.

The power fallback is the weaker argument that Hoenig & Heisey (2001) criticize. Among the 271 reference labels it decides six, all UNDERPOWERED, so no reference is REFUTED on power alone. Thirteen significant references skip the size test because their effect shares no scale with δ.

The deployed search. The search evaluates g_p, not f: significance at the label rule's α, and the direction while significant. It shifts the readings in the analysis table by $d\sigma$ between the two sides of the comparison and re-runs the same deterministic backend. The gates and the planned power at δ do not move with d: the gates read the stored measurement records, and the planned power depends only on sample sizes, event counts and, for the equivalence family, the SD of the paired differences. The grid runs from 0.05 to 1.00 in steps of 0.05, then 1.25, 1.5, 2.0, 2.5 and 3.0. At each point both orientations are tried, attenuating first, and the search returns the first point at which g_p changes. A conclusion that does not change by 3.0 is censored at the cap, not given an infinite margin. Both orientations are searched because measurement error need not bias an association toward the null (Yland et al., 2022).

The error scale. For centered errors e_A, e_B, $\mathrm{Var}(e_A - e_B) = \mathrm{Var}(e_A) + \mathrm{Var}(e_B) - 2\mathrm{Cov}(e_A, e_B)$. Only under zero covariance does disagreement identify the sum of error variances; it does not identify shared bias. The deployed scale assumes each instrument contributes between one third and two thirds of that variance, or 0.82–1.15 times the equal-share standard deviation. The check uses both endpoints; because a larger scale expresses the same shift in fewer error units, a run is flagged at either endpoint exactly when it is flagged at the upper one. Three pairwise variances identify three individual variances under uncorrelated errors (Grubbs, 1948), an assumption challenged by Appendix H.1. When three instruments read the same subjects and the estimate is positive, this difference form gives the run's instrument its own error scale, in which the counterfactual is then expressed. With two instruments, Grubbs's estimate $\mathrm{Var}(x_A) - \mathrm{Cov}(x_A, x_B)$ assumes a common measurement scale, and on 30 to 400 subjects it is a small difference of two large moments: it came out negative in 8 of 12 comparisons on our data. The pairwise disagreement is therefore split between the two allocation endpoints, a stated assumption rather than an estimate. Neither approach corrects shared error, which lowers every pairwise statistic (Section 6).

The four checks. Each check reads the run and two other repeats of the same hypothesis, and nothing else:

- *instrument*: the counterfactual is at or below one error unit under either allocation endpoint, the instruments reach different labels, or more than 5% of masks are empty for a structure present in every subject (the empty-mask gate's condition, which also makes the label INVALID);
- *label conflict*: the verdict differs from the verdict the run's own label implies, or contradicts the run's p-value, as when it answers YES after a non-significant result or NO after a significant one;
- *thin margin*: on δ's scale, the estimate or either interval bound lies less than one interval half-width from δ;
- *repeat disagreement*: at least one of the two repeats, chosen deterministically from the run's identifier, reaches a different verdict.

Unavailable checks contribute zero: missing intervals, secondary instruments, or sufficient repeats can therefore lower the score. The checks read no reference answer, enforced by a code-level test. The risk level changes neither label nor verdict.

B.2 PROOFS

Proof of Proposition 1. If $L = L_{\mathrm{ref}}$ then $m(L) = m(L_{\mathrm{ref}})$, so $v \neq m(L_{\mathrm{ref}})$ holds exactly when $v \neq m(L)$. If $L \neq L_{\mathrm{ref}}$, both sides of the equivalence hold. The event $v \neq m(L)$ reads the run's verdict and its own label and nothing else. □

The deployed label-conflict check flags this event. It also flags three warning codes that compare the verdict with the run's own p-value, which is why a few of the runs it flags are right on both counts (9 of 252 with the primary instruments; Appendix E.7).

B.3 WHAT THE SEARCH AND THE RISK LEVEL GUARANTEE

The search. Write $h(d) = g(S(\phi, T_d(D)))$ for one shift orientation. If d_j is the first grid point with $h(d_j) \neq h(0)$, then $d_g^* \leq d_j$ by definition, so the reported counterfactual is an upper bound on the decision-changing perturbation, and the smaller of the two orientations' values bounds the smaller threshold. The preceding point d_{j-1} is a lower bound only if the decision cannot leave its initial value and return to it between grid points. The implementation guarantees deterministic

statistics and fixed gates but does not establish this condition: a decision assembled from several thresholds can change and change back between two grid points. The search therefore does not certify stability below the reported value. A direction that never changes within the cap is censored at the cap rather than given an infinite margin, and a bound for g_p need not hold for the full label f (Sec. 5.5).

The risk level. If the threshold k were fixed independently of calibration outcomes and runs were independent and identically distributed, the number of wrong runs among those retained would be binomial, and a one-sided Clopper–Pearson limit would bound the error rate of future retained runs. Dependent repeats, related hypothesis templates, and a threshold chosen after inspecting the design corpus violate these assumptions, so the operating points in Appendix E.5 are empirical.

C SYSTEM DETAILS

C.1 PHASES AND ROLES

Saved outputs connect four phases. A principal investigator, imaging specialist, and statistician contribute to planning and interpretation through phase-specific prompts. Coding agents handle imaging requests and analysis; a critic provides bounded correction rounds. The following details specify the responsibilities outlined in Sec. 3.

- **Planning.** Agents specify structures, observations, population, test family, and covariates, or declare the hypothesis untestable. Mechanical checks reverse two kinds of refusal: one whose only reason is low power, because the label rule already reports an underpowered null as UNDERPOWERED, and one that calls a structure missing although the registered store holds it. Both decide on recorded facts, so a refusal for missing data or a missing modality stands. The run configuration fixes the imaging instrument; the plan may override it, but no reported run did.
- **Image analysis.** An agent requests registered subject, structure, and observation identifiers. Framework code executes the imaging tools (SAT (Zhao et al., 2025b), in its Pro version, VoxTell (Rokuss et al., 2026), and BiomedParse (Zhao et al., 2025a) for MRI; RETFound (Zhou et al., 2023) and VascX (Vargas Quiros et al., 2025) for fundus images), stores masks and readings, and records plausibility assessments.
- **Statistical analysis.** The framework builds a base table from stored measurements. Sandboxed agent code assembles the analysis table and typed specification. Framework code materializes declared columns, applies cohort selection, validates inputs, and computes statistics. Syntax-tree checks restrict dynamic execution, process and network access, and writes outside the workspace.
- **Interpretation.** Agents receive statistics and measurement information with the evidence label withheld, and write a verdict, explanation, and stated confidence.

C.2 THE TYPED ANALYSIS CONTRACT

`AnalysisSpec` is a closed schema specifying family, variables, grouping, predictors, covariates, direction, and α, with additional fields for paired, equivalence, and survival analyses. Eleven family names route to ten backends: Mann–Whitney, Wilcoxon signed-rank, paired equivalence test (TOST), Spearman correlation, Kruskal–Wallis, linear/logistic regression, chi-square/Fisher, log-rank/Cox, Jonckheere–Terpstra, and exact binomial tests. Significance thresholds may be stricter than 0.05 but never looser. Claimed direction is separate from the direction estimated from data.

`CohortSpec` defines eligible rows using recorded fields, the registered split, and aggregation level. It supports equality, membership, inequalities, and ranges, with no sampling or row-index operator. The framework applies these rules to the agent's table before statistical computation and preserves the input table for audit.

Effects are compared with δ on declared scales: pooled-SD Cohen's d for a Mann–Whitney contrast, d_z for paired contrasts, Spearman's ρ, standardized regression coefficients, and log hazard ratios per predictor SD. Non-comparable effects carry an explicit reason and skip the size test. Missing SDs prevent conversions of raw regression slopes or continuous hazard ratios. Missing or non-finite statistics, degenerate fits, violated test constraints, and failed plausibility gates yield INVALID; Appendix B defines the remaining label branches.

Table 3: **The three plausibility gates.** Each makes the evidence label INVALID; nothing else in the measurement layer can change a label.

Gate	Question	Rule	Threshold
Empty mask	Does a mask exist?	Share of stored masks with no positive voxel, per (structure, observation), for a plan-bound structure that is present in every subject.	5%
Volume floor	Is the number possible?	Share of measured masks below a literature-anchored volume floor for that structure.	5.0 mL floor, 5% of the cohort
Timepoint order	Are two numbers on one patient consistent?	Share of subjects whose end-diastolic ventricular volume is below their end-systolic one, a violation of the definition of the two timepoints rather than of a reference range.	5% of the cohort

C.3 THE SEVEN INTEGRITY CHECKS

The implementation and offline tests enforce seven requirements:

1. A deterministic backend computes statistics from the typed specification.
2. Result files are regenerated from specification and table; a missing specification fails the phase without a free-form fallback.
3. Saved inputs reproduce the statistics at a fixed code version.
4. Coverage checks and code scans detect cohort subsampling.
5. Code scans detect synthetic-data generation.
6. Code scans detect reads outside registered stores.
7. The interpreting model never receives the evidence label or scores its answer.

An auditor regenerates each run's statistics from its saved specification and table and classifies the result as agreement, a mismatch at the same code version, expected drift across code versions, a difference in table shape only, or unverifiable. The permutation test for ordered trends (Jonckheere–Terpstra) uses a fixed seed. In the design corpus, synthetic-data scans invalidate 0–1.9% of runs per cohort–instrument pair, and unsupported significance claims occur in 0–0.6%. Only infrastructure failures, including GPU out-of-memory errors, may be retried; agent failures remain in the attempted-run record.

C.4 THE MEASUREMENT LAYER

Secondary instruments. The same specification runs on each instrument's reading of a quantity. Columns from the run's own instrument supply the reported statistics; secondary readings use separate columns. A secondary instrument covering less than 80% of the rows of the run's instrument is unavailable. Comparisons use common rows, rerunning the analysis of the run's instrument there without changing its full-cohort statistical results. Agreement requires at least two comparable labels. Interpretation receives the number of completed instrument analyses and the number agreeing, without their labels.

Plausibility gates. Gates apply only to structures read by the executed specification, expected in every subject, and covered by a registered threshold (Table 3). Empty necrosis masks are therefore allowed: experts record no necrotic core in 89% of grade 2 tumors. Each rule encodes domain knowledge rather than a fit to outcomes: which structures every subject has, published lower limits for a volume, and the definition of the cardiac timepoints. Clinicians could supply such rules for a new cohort, or an agent could propose them for review.

The 5.0 mL ventricular floor is an impossibility threshold rather than a reference limit: it lies well below every published adult lower limit of normal (Kawel-Boehm et al., 2020; Petersen et al., 2017), and it excludes none of the clean cardiac measurements, whose smallest end-systolic volume is 12.4 mL. Anatomical-validity checks set their thresholds the same way, so that no valid expert segmentation is flagged, although on shape and topology rather than volume (Painchaud et al., 2020).

Table 4: **Benchmark composition.** Bank revision 2.2 (279 items). Tiers: L0 untestable on this cohort, L1 metadata only, L2 one segmentation-derived measurement, L3 engineered imaging features, L4 metadata and imaging combined, L5 multivariable or survival with covariates. Control types: *pos* positive control, *neg* negative control (the claim is expected to be refuted), *no-eff* no-effect control, *hard* hard negative, *under* underpowered control, *nons* nonsense control, *untest* untestable control. Tier L0 holds the five original untestable controls; the three PAPILA untestable controls keep the tier of the analysis they would need (optical coherence tomography, blood pressure, or longitudinal data).

	tier						**control type**						
cohort table	L0	L1	L2	L3	L4	L5	pos	neg	no-eff	hard	under	nons	untest
ACDC	3	5	46	5	5	0	36	13	4	5	3	0	3
M&Ms-2	0	2	55	2	0	0	32	10	4	9	4	0	0
UCSF-PDGM	2	14	12	7	21	6	34	10	6	3	5	2	2
PAPILA (eyes)	0	0	22	3	1	1	19	2	0	0	3	0	3
PAPILA (patients)	0	3	55	9	0	0	38	10	10	7	2	0	0
total	5	24	190	26	27	7	159	45	24	24	17	2	8

The timepoint gate applies to ventricles only: myocardium thickens during systole and atria follow a different phase. We omitted a rule that myocardial mass stays constant between the two timepoints: its median change is 7.8–9.2% whether the instrument performs well or poorly, so no tolerance separates failed measurement from normal variation.

The three gates share one cohort fraction, 5%, first set for empty masks, where it separated structures a tool can segment from one it cannot; a single fraction avoids tuning each gate. A fraction rather than a single failure keeps one unusable mask from voiding an analysis. With well-performing instruments, the volume floor and the timepoint order are never violated on ventricular structures.

Out-of-range fractions and ratios are recorded without gating: on glioma data their precision is below the wrong-label base rate. Instrument disagreement and the counterfactual also remain assessments rather than label-changing gates. Appendix B defines the search.

C.5 The measurement record

The measurement record links each available assessment to the instrument and step that produced it. Planning adds the planned power; imaging adds mask and timepoint checks; statistics adds instrument agreement, the assumed error scale, and the counterfactual. Interpretation records stated confidence separately. Entries cannot be rewritten or deleted, and the label rule does not read agent confidence. An unavailable scale or check supplies no assessment; the record does not attach a calibrated uncertainty distribution to every reading.

Interpretation receives two factual blocks: measurement reliability (error scale, offset, and whether that error changes the result) and instrument agreement (numbers compared and agreeing). The reliability block states: "This states how much measurement error the result tolerates. It does not state whether the hypothesis is true." The layer ablation removes both blocks without changing the agenda.

D Benchmark details

D.1 Cohorts and test families

ACDC contains cine cardiac MRI from 150 patients in five disease groups (Bernard et al., 2018); M&Ms-2, the successor to M&Ms (Campello et al., 2021), contains 360 subjects across vendors, centers, and eight groups (Martín-Isla et al., 2023). UCSF-PDGM supplies multi-sequence glioma MRI with survival and molecular data for 501 scans (Calabrese et al., 2022b) from 495 unique patients (Calabrese et al., 2022a). PAPILA provides 488 eyes from 244 patients, two human graders, and intraocular pressure from pneumatic and Perkins tonometry (Kovalyk et al., 2022). Its eye- and patient-level tables define different analysis units and reference answers.

Retinal models. Twenty-two PAPILA hypotheses, 11 per table, use RETFound and VascX outputs. RETFound (Zhou et al., 2023) was pre-trained on 1.6 million retinal images, 0.9 million of them color fundus photographs; the checkpoint used here was fine-tuned on PAPILA's training split to classify healthy, suspect, and glaucoma eyes. VascX (Vargas Quiros et al., 2025) uses public weights to segment vessels, arteries, veins, and the optic disc; the framework derives vessel density, artery-to-vein ratio, and pixel counts. Both models were run once on the 79 validation eyes, outside classifier training. All items using these outputs are restricted to those eyes; patient-level items average over available validation eyes. Per-eye runs request readings through the imaging interface, whereas patient tables contain precomputed columns. References use the same model outputs because no expert counterpart exists for these quantities. These items test analysis and interpretation. With no secondary instrument, their instrument check never flags any of the 220 runs; the other three checks determine their risk levels.

Each hypothesis declares family, direction, population, control type, and complexity tier, with frozen text and metadata hashes. The bank includes 150 two-group differences, 47 correlations, 24 survival, 21 paired differences, 10 adjusted regressions, 9 categorical associations, 9 equivalence, 8 trends, and one proportion test.

D.2 Construction and reference answers

New items were written from templates in four waves: 64 ACDC/UCSF-PDGM items, 57 M&Ms-2 items, 62 PAPILA-patient items, and 13 items built around known instrument failures. Code rendered each item's text from its tuple of cohort, family, variables, population, and covariates and a fixed glossary, without values or statistics. Directions came from cited priors or seeded coin flips for mechanical controls. Text and metadata were frozen before computing references.

Each item's registered analysis then ran through the live system's deterministic backend on expert-derived tables. The same label rule generated reference labels. After references existed, items could be excluded with a recorded reason but not edited. Some exclusions depended on reference instability, so freezing text did not eliminate reference-dependent selection.

The remaining 83 items predate these waves and were re-derived with typed analyses: 61 of the VERITAS benchmark's 64 items and 22 earlier PAPILA items. The earlier references often lacked spread estimates needed for the effect-size test. Re-derivation changed six reference labels. Only the eight untestable controls have reference labels not derived from statistics.

D.3 The smallest effect of interest

The threshold δ determines whether an effect is meaningful under the benchmark's rule. Large conventional effects can exclude established associations: the glioma cohort's survival hazard ratio for methylation of the MGMT (O^6-methylguanine-DNA methyltransferase) promoter is 0.69 [0.52, 0.93], consistent with the published prognostic direction (Hegi et al., 2005), but the earlier hazard-ratio threshold of 2.0 yielded REFUTED.

The revised procedure selects an absolute difference, preferring a published minimal important difference, then a guideline or trial-design criterion, then reference-method repeatability. Among defensible anchors it uses the smallest at or above measurement resolution, and converts to the rule's scale using the analyzed spread. Unanchored items use $d = 0.2$ and are flagged. Table 5 records the anchors; the frozen metadata record each item's source.

The anchors require interpretation. Grothues et al. (2002) define the coefficient of variability as the SD of paired scan–rescan differences divided by the mean, so 1.96 times the coefficient times the reference mean is an absolute repeatability coefficient; it is applied to between-group differences. RANO's within-patient progression criterion is applied to between-group differences using the volumes each item compares, including necrotic and edema volumes that RANO does not assess. The glaucoma anchors transfer in the same way: Goldmann repeatability is applied to pneumatic readings, and a within-eye visual-field change to between-group differences. None of these is a direct clinical validation of the benchmark threshold. Two reference decisions lie close to their boundaries, by 0.03 on a log hazard-ratio scale and 0.001 for correlation; no decision buffer is applied. Appendix E.10 tests threshold sensitivity.

Table 5: **Anchors of the smallest effect of interest.** Cardiac values are interstudy repeatability coefficients against adult reference means (papillary muscles counted in the left-ventricular mass and in the right-ventricular cavity), taken at the low end of the published between-group range. Glioma volume thresholds are relative to the volumes each item analyses. Items without an anchor use Cohen's $d = 0.2$ and are flagged as unanchored in the bank.

quantity	δ	**anchor**
left-ventricular end-diastolic volume (LVEDV), ejection fraction (LVEF), myocardial mass	7.9 mL, 3.1 points, 5.6 g	repeatability (Grothues et al., 2002) × reference mean (Kawel-Boehm et al., 2020)
right-ventricular end-diastolic volume (RVEDV)	11.9 mL	repeatability (Grothues et al., 2004) × reference mean (Kawel-Boehm et al., 2020)
glioma volumes	40% of the analyzed mean volume	RANO 2.0 (Response Assessment in Neuro-Oncology) volumetric progression ($\geq 40\%$ over nadir, measurable tumor) (Wen et al., 2023)
survival	hazard ratio 1.25	phase-3 glioma trial design (Abo Kasem et al., 2026)
vertical cup-to-disc ratio	0.093; 0.066 ($0.093/\sqrt{2}$) for the two-grader mean	agreement of the two PAPILA graders (1.96× SD of their difference); inter-expert bound (Varma et al., 1992)
intraocular pressure	2.2 mmHg	Goldmann applanation repeatability (Kotecha et al., 2010)
visual-field mean defect	4 dB	change detectable over two years with three fields per year (Chauhan et al., 2008)
no anchor (e.g. RETFound probabilities, vessel density, age)	$d = 0.2$, flagged	convention

Dropped and held-out items. Thirty-six items were excluded for recorded reasons: inconsistent text and metadata, duplication, no single expressible binding, small groups or unstable tonometer comparisons, inappropriate nulls, or dependence on units or grader choice. Eight further items remain in the bank but are held out of the accuracy pools: seven have conflicting defensible reference readings, and one changes label between patient- and eye-level analysis. Exclusions and family-specific eligibility explain why accuracy denominators differ from the bank size.

D.4 Specification agreement

Each hypothesis has a registered analysis: the test family, variables, groups, covariates, and direction used to compute its reference answer. For every run, we compare the analysis the agents specified with this registered one, field by field. Eleven fields are compared, but only those the chosen test uses; groups and covariates are compared as sets. The comparison needs the reference, so it is an offline diagnostic and enters neither the checks nor the scoring.

A field can match, differ, or be impossible to compare. The last case arises when the registered analysis reads a column of the cohort table while the agents measured the same quantity themselves under another name; without a mapping between the two names, we count neither a match nor a mismatch.

Runs matching the registered analysis have 98%, 94%, and 98% label accuracy on ACDC and the two PAPILA tables. Different analyses can also produce the reference label: $13/48$ runs with a different test family do so on those tables. Conversely, matching the analysis is insufficient when measurements are inaccurate: accuracy is 36% and 16% on M&Ms-2 with SAT and VoxTell. All 148 runs with no specification are wrong.

The comparison is incomplete: many cardiac outcome names cannot be compared, whereas PAPILA runs and references share column names. The most common mismatch is the declared direction (709 of $4{,}942$ comparable rows). It rarely changes the label, because the label rule compares the observed

Table 6: **Specification agreement and label correctness.** Runs with a correct evidence label / runs in each bucket (percent); gates on, scored rows. Buckets compare the run's analysis with the item's registered analysis: *same analysis*, every compared field agrees; *names not comparable*, no field disagrees, but the run measured its own quantity under its own name with no mapping to the registered cohort column; *different variables*, same test family but another field differs; *different family*, a different test family; *no spec*, the run registered no analysis.

cohort	same analysis	names not comp.	different variables	different family	no spec
ACDC	398/405 (98)	86/94 (91)	78/98 (80)	1/3 (33)	0/10 (0)
UCSF-PDGM (VoxTell)	212/236 (90)	69/86 (80)	180/243 (74)	0/13 (0)	0/22 (0)
UCSF-PDGM (SAT)	173/229 (76)	49/79 (62)	199/252 (79)	0/21 (0)	0/19 (0)
UCSF-PDGM (BiomedParse)	212/234 (91)	77/88 (88)	182/249 (73)	1/13 (8)	0/16 (0)
PAPILA (eyes)	82/87 (94)	5/5 (100)	83/108 (77)	10/36 (28)	0/4 (0)
PAPILA (patients)	285/290 (98)	26/27 (96)	198/253 (78)	2/9 (22)	0/11 (0)
M&Ms-2 (BiomedParse)	344/391 (88)	34/55 (62)	85/116 (73)	0/7 (0)	0/21 (0)
M&Ms-2 (SAT)	138/382 (36)	28/60 (47)	26/113 (23)	0/17 (0)	0/18 (0)
M&Ms-2 (VoxTell)	63/403 (16)	5/59 (8)	15/88 (17)	1/13 (8)	0/27 (0)

Table 7: **Pools behind each rate.** Held-out items are excluded from every pool except attempted runs.

pool	definition	used for
attempted runs	every run started, including agent failures and refusals	coverage over all attempts
completed runs	runs that pass all four phases	completion
answerable rows	runs with a reference label and a computed label, excluding the eight untestable controls; a gate refusal counts as a wrong label	label accuracy
parsable verdicts	runs whose verdict could be extracted	verdict accuracy
scorable runs	runs whose verdict and label the evaluator can judge	error shape (Table 13)
eligible runs	scorable runs on items with a non-INVALID reference label whose own label is not INVALID	risk levels, selection, AURC

direction with the direction in the hypothesis, not with the agents' declaration. A correct label alone therefore does not show that the intended question was tested.

E ADDITIONAL EXPERIMENTS

This appendix follows the questions of Section 5: system accuracy, selection and transfer, instruments and individual checks, the measurement-layer ablation, and the counterfactual under injected error. A final subsection tests how the evaluation depends on δ. Table 7 defines the pools of runs and items behind each rate; combined error means a wrong label or verdict. AURC, the area under the risk–coverage curve, measures how well a score orders runs from least to most likely wrong: lower is better, an oracle attains the minimum, and an uninformative ordering has AURC equal to the pool's error rate.

E.1 SYSTEM ACCURACY

The Qwen3.8 team's verdict advantage over the single-model agents also holds on all available items (Table 8).

The gpt-oss team uses `gpt-oss:20b` (OpenAI, 2025) as lead and imaging agent, `qwen3:8b` (Yang et al., 2025) as statistician and critic, and `qwen3-coder:30b` (Qwen Team, 2025) for code, with 16k context and temperature 0.2. All three baselines use `qwen3.8:27b` (Qwen Team, 2026), ten repeats, the same bank, context limits, temperature, and integrity scans. SM-Code writes code, results, and intervals with one attempt and two correction rounds. SM-Typed fills a typed specification without correction; SM-Typed+2 adds two rounds of framework error feedback. VERITAS-M also changes phases, information flow, and critic use. Table 1 restricts comparisons to hypotheses run by every system in its block: the Qwen3.8 team did not run the last nine PAPILA-eye and two M&Ms-2 items. Table 8 retains all available items. All sys-

Table 8: **Accuracy on available items.** Label / verdict accuracy (%), ten repeats per item. VERITAS uses the 30 ACDC and 31 UCSF-PDGM items kept from its benchmark, which has no M&Ms-2 or PAPILA items. Its label accuracy is not comparable (—): only 2.4% of its runs have a usable effect interval, which the label rule needs. Its verdict accuracies re-score runs of the public VERITAS code under the current bank and evaluator, so they differ from those reported by Stoffl et al. (2026). Other systems use 64, 62, 59, 27, and 67 items, respectively, except the Qwen3.8 team (64, 62, 57, 18, 67). *No measurement layer* is the ablation in Table 11; the last row substitutes SAT for the primary instrument on M&Ms-2 only. Accuracy pools exclude held-out items (Appendix D); verdict accuracy also excludes runs without a parsable verdict. Differences between rows therefore need not equal those in Table 1, which uses only items every system ran. SM-Code, SM-Typed, and SM-Typed+2 are single-model agents (Sec. 5.1); the Qwen3.8 team uses `qwen3.8:27b` in its main roles and `gemma4:31b` as critic (Gemma Team, 2026).

system	models	ACDC	UCSF-PDGM	M&Ms-2	PAPILA (eyes)	PAPILA (patients)
Items kept from the VERITAS benchmark						
VERITAS (Stoffl et al., 2026)	gpt-oss team	— / 81.6	— / 48.1	*not in the VERITAS benchmark*		
VERITAS-M	gpt-oss team	93.0 / 93.0	76.2 / 71.9			
All items						
SM-Code	Qwen3.8	89.2 / 83.1	38.3 / 68.2	78.6 / 67.3	36.7 / 70.0	73.6 / 61.6
SM-Typed	Qwen3.8	98.4 / 79.8	51.0 / 54.5	78.5 / 70.1	78.8 / 78.1	88.1 / 66.7
SM-Typed+2	Qwen3.8	99.8 / 81.5	83.7 / 66.9	83.1 / 72.4	87.5 / 85.9	91.5 / 70.3
VERITAS-M, no measurement layer	gpt-oss team	92.1 / 94.2	76.8 / 71.5	78.6 / 85.8	80.8 / 82.3	86.9 / 85.6
VERITAS-M	gpt-oss team	92.3 / 95.3	76.8 / 69.9	78.5 / 83.5	75.0 / 81.1	86.6 / 84.6
VERITAS-M	Qwen3.8 team	97.2 / 99.1	85.0 / 83.8	84.7 / 84.9	96.0 / 98.9	90.8 / 96.0
VERITAS-M, SAT substituted	Qwen3.8 team	*not run*		41.4 / 65.1	*not run*	

tems share the evaluation interface; evaluating another system's evidence labels requires statistical results in a compatible format.

With equal correction rounds, typed specifications improve evidence labels on all five cohort tables (SM-Typed+2 versus SM-Code), and corrections increase UCSF-PDGM completion from 54% to 88%. SM-Typed+2 has higher label accuracy than the gpt-oss team on all five tables, whereas the gpt-oss team has higher verdict accuracy on four tables in Table 1. VERITAS-M's advantage with the Qwen3.8 team, which shares the baselines' primary model, is larger for verdicts than for labels; its additional critic uses another model family.

Layer-on versus layer-off gpt-oss-team rows measure a different effect (Sec. 5.4; Appendix E.8). SM-Code has no compatible measurement record, so its risk level uses three checks.

In the design corpus (5,210 gpt-oss-team runs over nine pairs), 81 to 94% of runs complete all four phases in each pair; label accuracy per pair appears in Table 11. Untestable controls are correctly declined on 23/30 ACDC runs, 14–18/20 UCSF-PDGM runs, and 14/30 PAPILA-eye runs.

E.2 Comparison with the earlier system

Re-scoring runs of the public code of VERITAS (Stoffl et al., 2026) under the current evaluator changes its published scores only slightly, so the comparison in Section 5.1 does not depend on the change of evaluator. One evaluator version scores every system and re-derives the statistical results of runs and references before comparing them, so older results without spread estimates do not receive a weaker label rule. We re-scored 12,355 runs of VERITAS's public code (34 sets) under the current evaluator. Published scores change by a median -2.6 points, and the ranking of twelve configurations keeps a Spearman correlation of 0.94 with the published one. Only 2.4% of these runs' results have a usable effect interval, so their evidence labels are not comparable. On the 30 ACDC and 31 UCSF-PDGM items kept from its benchmark, VERITAS yields parsable verdicts on 267/300 and 208/310 runs and VERITAS-M on 256/300 and 256/310. Attempted-run gains are 6.7 points [0.0, 14.3] and 27.1 [15.2, 38.4]; majority verdicts are correct on 24 versus 27 and 17 versus 23 items.

E.3 Single-call interpretation

Replacing the interpretation meeting with a single model call preserves almost every verdict, so the meeting alone does not explain VERITAS-M's advantage. We copied the Qwen3.8 team's runs,

preserving plans, masks, specifications, statistics, and measurement records. Only interpretation changed: one call replaced the meeting, using the lead's prompt, the same agenda, and summaries of earlier phases. The standard extraction, validation, and scoring code processed its verdict. Runs without statistical results were copied unchanged.

The replacement preserves $2{,}466$ of $2{,}526$ re-interpreted verdicts (97.6%). Verdict accuracy moves from 99.1 to 99.5% on ACDC, 83.8 to 84.5% on UCSF-PDGM, and 96.0 to 96.2% on PAPILA patients, and is unchanged on M&Ms-2 and PAPILA eyes. On PAPILA patients, level 0 contains $5/168$ wrong runs with one call and $4/167$ with the meeting, all from a single reversed-group-coding item. The experiment does not isolate which inputs explain the advantage over SM-Typed+2.

E.4 Selection by risk level

In every cohort–instrument pair of the design corpus, no run at level 0 is wrong ($0/859$ over nine pairs), and pooled error rises with every further level (Table 11). The ordering also holds when repeats are aggregated to hypotheses and, as shown below, when the check thresholds change.

Each hypothesis receives majority labels and verdicts over eligible repeats; an item is wrong if either majority disagrees with the reference. Ties use a fixed order (Supported, Refuted, Underpowered; Yes, No, Inconclusive), affecting one label and ten verdicts among the 262 primary-instrument items. The risk level is the mode among runs supporting the majority label, with ties resolved upward (14 items). Taking the median or the minimum level over all eligible repeats instead gives AURC 0.035 and 0.034. The 651 level-0 runs with primary instruments come from 109 hypotheses; treating hypotheses as independent units, the one-sided 95% Clopper–Pearson bound on the share of hypotheses with a wrong level-0 run is 2.7%.

With primary instruments, selection retains $159/262$ items with two errors: 60.7% coverage [54.6, 66.8] at 1.3% error [0.0, 3.2]. Label-only error is zero among retained items. AURC is 0.0295 [0.0167, 0.0451]; its paired difference from the run-level result is -0.0011 [-0.0071, 0.0057]. For the transfer systems, selection keeps 151, 154, and 177 items with 6, 10, and 3 wrong for SM-Typed, SM-Code, and the Qwen3.8 team; the Qwen3.8 team's one wrong item at level 0 is the PAPILA hypothesis with reversed group coding.

Agent confidence is weaker on the same $2{,}308$ eligible primary-instrument runs. High-confidence runs have 9.7% error over $1{,}349$ runs, medium 22.4% over 603, and low 24.7% over 356. Its AURC exceeds the risk level's by 0.080 [0.058, 0.104]. The curves handle tied scores without arbitrary within-bucket ordering.

Sweeping the thin-margin threshold from 0.5 to 2 interval half-widths changes AURC by 0.0032 and leaves level 0 without observed errors. Varying how repeat disagreement is counted changes it by 0.0102; the strictest variant admits at most $5/852$ wrong runs at level 0, and removing the repeat check leaves 11 wrong runs from 10 hypotheses at level 0. No variant improves AURC with a paired interval excluding zero. The deployed thresholds were fixed in natural units after inspecting the design corpus and before the per-level error rates were computed.

E.5 Transfer and offline score variants

The fixed rule leaves level 0 nearly free of errors on all three systems scored after it was fixed ($0/755$, $2/827$, and $4/820$ wrong runs; Fig. 3b). At $r < 2$, SM-Typed retains 64% of eligible runs at 3.6% error, SM-Code 71% at 6.1%, and the Qwen3.8 team 71% at 1.2%. These systems reuse the benchmark, so they test transfer across systems rather than to new scientific domains or independently sampled hypotheses. Under label error alone, $r < 2$ keeps wrong labels at 0.7% for SM-Typed, 2.9% for SM-Code, and 0.4% for the Qwen3.8 team.

Two offline variants ask whether a rule fitted to outcomes would do better. Each is fitted on half of the hypotheses and evaluated on the other half, over 200 random splits with primary instruments. The first orders the checks by their precision on the fitting half. It lowers the mean test-half AURC from 0.030 for the deployed count (0.031 on the full pool) to 0.022, and fitted weights add little (0.021). The learned order puts label conflict first and the instrument check third or fourth, which reflects this pool: with primary instruments, consistency errors dominate. With substituted instruments, the

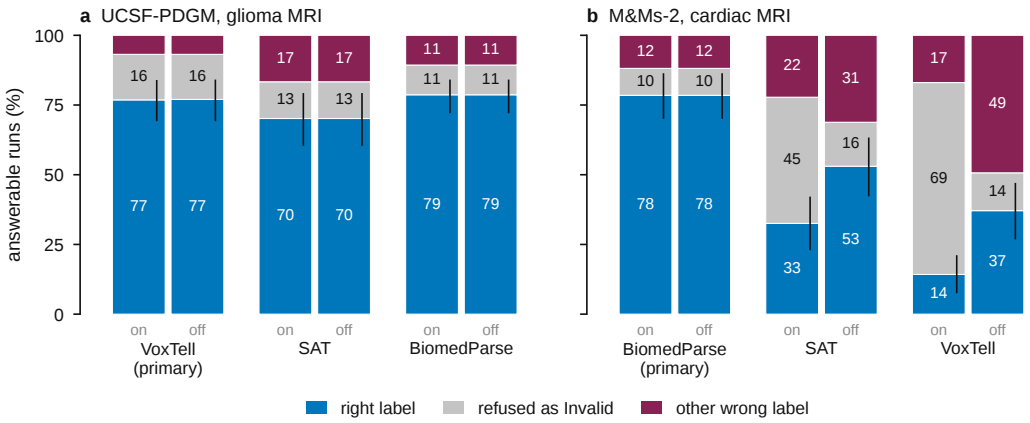

Figure 4: **Instrument choice.** Substituting the instrument changes both the answer and the refusal rate. The same hypotheses were run with each instrument supplying the readings (the cohort's primary instrument first) and scored with the plausibility gates on and, on the same runs, off. Each bar splits the answerable runs into a right label, a run refused as INVALID, by a gate or another invalid route, and any other wrong label; whiskers are 95% bootstrap intervals over hypotheses for the right share. On glioma MRI the three instruments are close (70 to 79% right, 11 to 16% refused). On cardiac MRI, SAT and VoxTell get 33% and 14% right, and 45% and 69% of their runs are refused, of which the gates account for about 30 and 55 points. With the gates off, 21 and 23 points more of their runs are right and 9 and 32 points more are wrong. On ACDC and the two PAPILA tables the gates change nothing.

instrument check carries most of the label-error ranking (Table 2), so we kept the count, which was fixed before these analyses.

The second variant chooses the loosest threshold whose one-sided binomial bound on the fitting half meets a target error. A 5% target selects $r < 2$ on 199/200 splits, with 61% coverage and 1.5% error. The same threshold keeps 3.6% error on SM-Typed, but 6.1% on SM-Code and 9.6% with substituted instruments. A 2% target selects level 0 and stays within 2% on all three sets; a 10% target keeps levels 0 to 2 and exceeds 10% on all three. Thresholds chosen on outcomes with dependent runs give empirical operating points, not certified bounds, and neither variant changes the deployed rule.

E.6 INSTRUMENT COMPARISONS AND THE NECROSIS EXAMPLE

The primary instruments were selected before runs measured under them, but reference use differed by cohort. ACDC keeps SAT, the segmentation tool of the VERITAS benchmark, and PAPILA's readings come from its two human graders. M&Ms-2's primary was chosen using ACDC contours: BiomedParse had the smallest error SD on each cardiac biomarker, including RVEDV (6.7 mL versus 10.2 for SAT and 15.3 for VoxTell). On M&Ms-2 its RVEDV bias is -0.7 mL versus -59.2 and -76.5 mL. For each primary, the other two tools serve as secondary instruments.

UCSF-PDGM's choice followed poor SAT performance on necrosis hypotheses and used expert masks from the evaluation cohort itself. We compared all three instruments on 495 subjects using contrast-enhanced T1 images (T1c), and recomputed necrosis contrasts on the same case set (Table 9). VoxTell was then designated primary. This evaluates the selected instrument but does not demonstrate reference-free selection or selection generalization to a new glioma cohort.

Experts record no necrotic core in 89.3% of grade 2 and 83.7% of grade 3 tumors. These are annotated absences: all 100 zero-necrosis masks contain tumor, and 86 are grade 2 or 3. SAT records zeros in only 17.9% and 7.0% of those grades, with median volumes 4.30 and 5.61 mL versus expert medians of zero. Grade 4 volumes are closer (7.70 versus 6.60 mL). Filling the low-grade end of the distribution reduces the grade contrast from 0.906 to 0.199 and reverses necrotic-fraction contrasts. BiomedParse instead misses necrosis in 18% of cases with expert-positive masks.

Table 9: **Instruments against expert contours, UCSF-PDGM.** Three instruments against the dataset's expert segmentations, 495 subjects, T1c. NEC: necrotic core; vol.: its volume; frac.: its share of the whole-tumor volume; 4 vs 2: WHO grade 4 against grade 2; IDH: isocitrate dehydrogenase status (wild type, wt, or mutant, mut). *Dice* is the median over cases where at least one of the two masks is non-empty; *empty* is the share of cases where the tool stored a mask with no positive voxel while the expert mask was not empty; ρ is the Spearman correlation between the tool's volume and the expert's. Contrast rows are Mann–Whitney rank-biserial effects computed once per source on one shared case set.

	necrotic core			**median Dice, other structures**		
tool	Dice	empty	ρ	whole tumor	enhancing	edema
SAT	0.691	0.25%	0.666	0.844	0.708	0.707
VoxTell	**0.896**	3.04%	**0.833**	**0.866**	**0.889**	**0.763**
BiomedParse	0.363	17.97%	0.687	0.762	0.794	0.712

Contrasts (rank-biserial), same 495 cases, once per source

	NEC vol., 4 vs 2	NEC vol., IDH-wt vs mut	NEC frac., 4 vs 2	NEC frac., wt vs mut
expert	**0.906**	**0.674**	**0.869**	**0.698**
SAT	0.199	−0.061	−0.160	−0.252
VoxTell	0.692	0.425	0.660	0.424
BiomedParse	0.686	0.450	0.683	0.470

A reference-free ranking from the tools' own masks, which scores each instrument by its rates of empty, implausibly small, and timepoint-inconsistent masks, ranks SAT first because SAT rarely returns an empty mask. For a structure that may be absent, such as necrosis in low-grade tumors, this rewards false positive readings, so reference-free instrument ranking cannot replace a measurement check. VoxTell improves the grade and IDH contrasts but recovers only 76% and 63% of the reference effects. Its estimates remain below the declared threshold on both IDH contrasts.

E.7 Error definitions and individual checks

At least one check flags every wrong label in the design corpus, and the instrument check contributes most when instruments are substituted. We evaluate the same eligible runs under combined, label-only, and verdict-only error. No run at level 0 is wrong under any of the three definitions in any cohort–instrument pair. Table 2 in the main text splits the checks by instrument group, and Figure 5 shows the check predicates and drop-one results. Label-only evaluation is essential because verdict–label contradictions are already errors by Proposition 1.

Among 252 label-conflict flags with primary instruments, 213 have a correct label and wrong verdict, 17 a wrong label and correct verdict, 13 both wrong, and 9 both right. The last group comes from additional p-value warnings. On M&Ms-2 with SAT and VoxTell, $31/41$ and $28/36$ flagged runs instead have wrong labels.

Components of the instrument check. The check combines counterfactual sensitivity, label disagreement, and the coverage condition; the last coincides with the empty-mask gate and never flagged an eligible run. With substituted instruments, label disagreement flags wrong labels with 70% precision and the counterfactual with 40%; for the Qwen3.8 team, the counterfactual flags 88% of wrong labels. Replacing the check with either active component reveals complementary behavior (Table 10). With substituted instruments, disagreement alone improves AURC relative to the full check, while counterfactual alone leaves eight wrong runs at level 0, from one hypothesis. For the Qwen3.8 team, removing the counterfactual raises retained label error from 0.4 to 1.2%. Removing the entire instrument check worsens label-error AURC in every displayed pool. These variants were evaluated after fixing the rule.

E.8 Layer ablation and gate-only rescoring

Without the measurement layer, the risk level keeps more runs at higher error and admits 20 wrong runs to level 0 (Table 11); on the same runs, the gates cause most of the accuracy differences on M&Ms-2. The full ablation removes secondary instruments, plausibility records, per-instrument

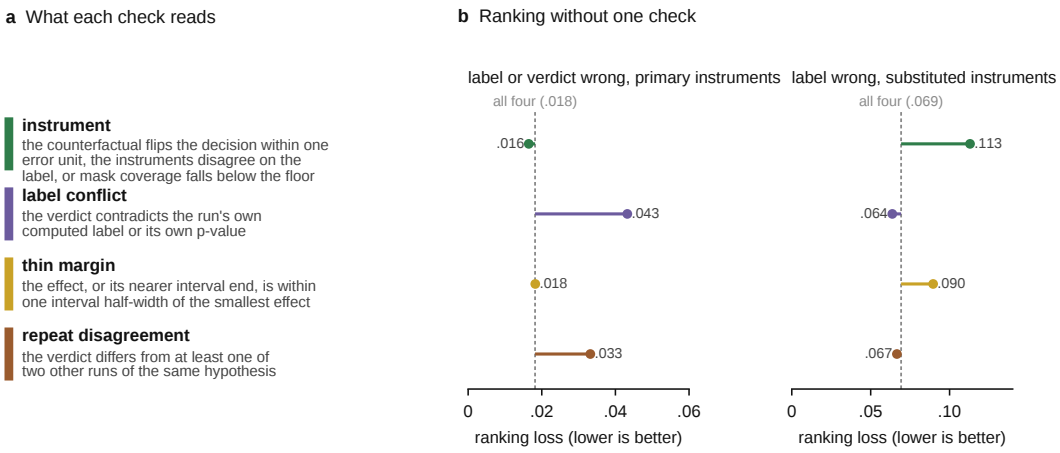

Figure 5: **The four checks.** What each check reads, and what the ranking loses without it. (a) The predicates as coded; the two thresholds are read from the rule. (b) Ranking loss, the area under the risk–coverage curve above a perfect ordering (Fig. 3a), when the rule is rebuilt without one check; a dot right of the dashed line (all four checks) means the ranking is worse without that check. Left: label or verdict wrong, each cohort's primary instrument (2,308 eligible runs); the label-conflict and repeat checks carry the ranking, and removing the instrument check slightly improves it. Right: label wrong, substituted instruments (1,557 eligible runs); removing the instrument check costs most, raising the loss from 0.069 to 0.113.

Table 10: **The two parts of the instrument check.** Label error only. The risk level is rebuilt with the instrument check complete (*full*), reduced to its counterfactual part (the decision changes within one error unit) or its disagreement part (instruments produce different labels), or removed (*none*); the other three checks are unchanged. AURC: area under the risk–coverage curve (lower is better). No wrong run reaches level 0 except with substituted instruments under the counterfactual alone (8 of 222, one hypothesis) or with no instrument check (21 of 351, three hypotheses), and for the Qwen3.8 team, whose four wrong runs at level 0 come from one hypothesis in every variant.

	AURC, label error				**keep $r < 2$: label error (%)**			
pool	full	counterfactual	disagreement	none	full	counterfactual	disagreement	none
design corpus, primary instruments	0.016	0.017	0.015	0.019	1.0	1.3	1.3	2.7
design corpus, substituted instruments	0.104	0.119	0.090	0.148	5.1	6.6	5.3	16.1
SM-Typed	0.007	0.009	0.007	0.010	0.7	1.4	0.8	1.7
Qwen3.8 team	0.011	0.011	0.012	0.019	0.4	0.4	1.2	4.0

Table 11: **Measurement-layer ablation.** gpt-oss team with the measurement layer (*on*) and without its gates, instrument comparisons, and measurement information for interpretation (*off*). Label and verdict accuracy (%); wrong/total eligible runs at level 0; coverage and retained combined error (verdict or label wrong) at $r < 2$, as fractions. Label accuracy uses answerable, non-held-out items. Accuracies are not pooled across cohort tables (—). Both conditions use ten repeats on different runs; eligibility can differ because the layer includes gates.

	label		**verdict**		**wrong at level 0**		**keep at < 2, on**		**keep at < 2, off**	
cohort, instrument	on	off	on	off	on	off	cover	error	cover	error
ACDC	92.3	92.1	95.3	94.2	0/293	0/303	0.81	0.00	0.87	0.00
UCSF-PDGM (VoxTell)	76.8	76.8	69.9	71.5	0/99	0/90	0.46	0.05	0.61	0.06
UCSF-PDGM (SAT)	70.2	69.7	64.7	66.3	0/99	13/137	0.47	0.11	0.61	0.17
UCSF-PDGM (BiomedParse)	78.7	77.2	75.9	73.8	0/87	0/123	0.47	0.09	0.61	0.08
M&Ms-2 (BiomedParse)	78.5	78.6	83.5	85.8	0/98	0/216	0.45	0.01	0.83	0.07
M&Ms-2 (SAT)	32.5	51.7	62.9	64.5	0/16	3/110	0.28	0.05	0.65	0.23
M&Ms-2 (VoxTell)	14.2	36.8	41.2	41.9	0/6	4/22	0.13	0.22	0.53	0.42
PAPILA (eyes)	75.0	80.8	81.1	82.3	0/42	0/61	0.77	0.03	0.80	0.02
PAPILA (patients)	86.6	86.9	84.6	85.6	0/119	0/136	0.64	0.01	0.64	0.01
pooled	—	—	—	—	0/859	20/1198	0.52	0.04	0.68	0.11

Table 12: **Matched coverage.** Measurement layer on against off (Table 11); combined error (verdict or label wrong). *on*: the layer-on runs' coverage and error at $r < 2$; *off at on's coverage*: the layer-off runs' error at that coverage, read off their risk–coverage curve with runs at the same level in random order (expected error), with a 95% bootstrap interval over hypotheses. The two conditions are different runs with different eligible pools.

cohort, instrument	on: cover / error (%)	off at on's coverage (%)	n on	n off
ACDC	80.6 / 0.0	0.2 [0.0, 1.0]	562	563
UCSF-PDGM (VoxTell)	45.5 / 5.3	5.1 [1.4, 10.7]	501	501
UCSF-PDGM (SAT)	47.2 / 10.6	15.4 [5.5, 27.7]	519	517
UCSF-PDGM (BiomedParse)	46.6 / 8.9	6.7 [1.3, 13.7]	532	517
M&Ms-2 (BiomedParse)	45.2 / 0.8	1.3 [0.0, 7.4]	526	525
M&Ms-2 (SAT)	28.2 / 5.5	8.6 [0.0, 23.6]	323	482
M&Ms-2 (VoxTell)	12.6 / 21.7	35.0 [9.7, 55.4]	183	507
PAPILA (eyes)	76.6 / 3.4	1.8 [0.0, 9.0]	192	199
PAPILA (patients)	63.8 / 0.6	1.5 [0.0, 6.8]	527	530

analyses, and both measurement blocks shown to interpretation. The run's own measurements, base-table construction, and the typed backend remain. It uses the same nine pairs, commit, bank, and ten-repeat design (5,210 runs), but different runs.

Gate-only rescoring instead compares the same runs with gates enabled or disabled (Fig. 4). The main changes occur on M&Ms-2 with SAT and VoxTell; other pairs change by no more than 0.2 accuracy points. With VoxTell, 69% of labels are INVALID, and the gates account for 55 of those points. Without them, 37% match the reference and 74% of REFUTED outputs are false refutations. Gates therefore suppress both wrong and some reference-matching answers based on implausible measurements.

Under label error alone, the design corpus without the layer keeps 9.8% wrong labels at $r < 2$. Table 12 compares the two layer conditions at matched coverage using tie-aware curves. The conditions contain different runs and eligible pools, so neither matching coverage nor dropping one check turns the layer ablation into a controlled intervention on the counterfactual alone.

E.9 COUNTERFACTUALS UNDER INJECTED ERROR

Under injected differential error, the counterfactual tracks the shift at which the full label changes but slightly overstates it, mainly because of the search grid. For $141/271$ typed items with a characterized continuous biomarker, we shift expert readings in both differential directions on a grid five times finer than the deployed search. We compare the reported counterfactual with the first shift

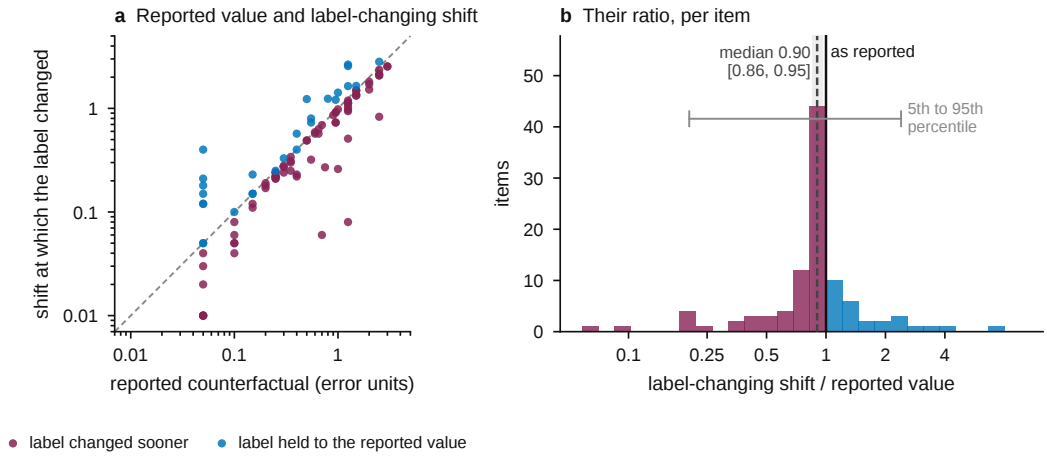

Figure 6: **Counterfactual against injected error.** How tight the reported counterfactual is, on the 102 items with both numbers. (a) The reported counterfactual against the shift at which the label changes under injected error, on log axes with the identity line dashed; red points (75 items) changed sooner than reported, blue ones held. (b) Their ratio, colored the same way. The median is 0.90 with a cluster-bootstrap interval of $[0.86, 0.95]$, entirely below one, so the reported counterfactual is not a conservative bound; the 5th to 95th percentile range, 0.20 to 2.40, is the per-item scatter.

changing the full label. Of these, 102 have both thresholds; the remaining 39 do not change within the searched cap, including 35 also censored by the counterfactual.

The ratio of the label-changing shift to the reported counterfactual has median 0.90 [0.86, 0.95] across 19 structure clusters. Regression gives slope 0.85 [0.67, 0.95], intercept 0.06 [0.01, 0.17], and $R^2 = 0.78$. The reported counterfactual is therefore not conservative: at the median, the label changes at 90% of it. Comparing deployed and fine-grid significance thresholds yields median ratio 0.91 and $R^2 = 0.99$, identifying discretization as the systematic component. Comparing fine-grid significance and full-label thresholds gives median ratio 1.00 and $R^2 = 0.78$: they agree on 60 items, the label changes first on 22, and significance first on 20. A finer grid reduces discretization error but does not reconcile the two decision functions, consistent with significance-based fragility tracking the p-value rather than effect size (Carter et al., 2017). Near the flag threshold of one error unit the deployed grid is fine (steps of 0.05), so discretization matters little for the instrument check; the difference between the significance decision and the full label matters more.

A second sweep varies assumed scale $a\sigma$ relative to injected bias $b\sigma$. At $a = b$ and injected biases near one error unit, the check detects 84–87% of label changes at 2–4% false alarms; at $a = 1.5b$, it detects 99% and flags about one fifth of unchanged labels; at $a = 0.5b$, it detects about two thirds. Of 127 observed label changes, 32 create rather than erase support, and 55 change SUPPORTED to UNDERPOWERED. Both shift directions and the full label therefore matter. The 130 out-of-scope items include 60 without a single continuous biomarker, 11 involving unmeasured metadata, and quantities without characterized error scales. Paired and equivalence designs were not injected (Appendix G).

E.10 SENSITIVITY TO THE SMALLEST EFFECT OF INTEREST

Selection changes little with δ when the checks read the same δ that defines the reference labels, but degrades when the two differ. We multiply each item's δ by c, recompute reference and run labels with gates fixed, and rebuild thin-margin and label-conflict flags. At $c = 1$, this reproduces all 271 reference labels and all 3,865 eligible runs' labels and levels. Verdicts remain those written under the original threshold, so the sweep assesses evaluation and selection sensitivity, not agent behavior under revised hypotheses.

With the checks recomputed at each scaled threshold (c from 0.5 to 2), label accuracy stays between 85.6 and 87.9%, level 0 holds at most 8 wrong runs and retained error stays between 3.6 and 8.5%, although 16 to 52 reference labels change. With the checks left as deployed while the references

Table 13: **Where the errors sit, per cohort and instrument.** Design corpus (gpt-oss team), ten repeats. Each scorable run is *right*, *verdict wrong only* (a correct evidence label misread, the case the label-conflict check targets), *label wrong only* (a wrong evidence label with a correct verdict), or wrong in *both*; percentages are of *scorable*. *Unfinished* counts runs without a judged verdict; they are not in *scorable*.

cohort, instrument	scorable	right	verdict wrong only	label wrong only	both	unfinished
ACDC	570	94%	4%	1%	1%	42
UCSF-PDGM (VoxTell)	544	66%	18%	3%	12%	58
UCSF-PDGM (SAT)	555	60%	16%	4%	21%	47
UCSF-PDGM (BiomedParse)	558	71%	13%	4%	12%	43
M&Ms-2 (BiomedParse)	534	82%	4%	1%	13%	56
M&Ms-2 (SAT)	510	36%	2%	27%	35%	80
M&Ms-2 (VoxTell)	536	14%	1%	27%	57%	54
PAPILA (eyes)	206	81%	6%	3%	11%	37
PAPILA (patients)	560	81%	11%	4%	5%	37

move, level 0 holds 70 and 86 wrong runs at $c = 1.5$ and 2. The thin-margin and label-conflict checks read δ, so a mismatch between the threshold the checks use and the one that defines correctness harms selection as well as accuracy.

F WHERE THE ERRORS COME FROM

Verdicts that misread a correct label dominate the errors on ACDC, UCSF-PDGM with VoxTell, and PAPILA patients; wrong labels dominate the other pairs, most strongly M&Ms-2 with SAT and VoxTell. Table 13 separates verdict-only errors, label-only errors, and runs where both are wrong. A correct verdict with a wrong label can arise because distinct labels map to the same verdict; it does not establish that the agent detected a measurement error. Unfinished runs are listed separately, using the denominators of Sec. 4.

Interpretation errors dominate UCSF-PDGM with VoxTell and PAPILA patients. Many answer NO to an UNDERPOWERED label (62 and 22 runs); 12 ACDC runs do the same. Conversely, 27 PAPILA-patient and 23 UCSF-PDGM runs answer INCONCLUSIVE where the reference says NO. On M&Ms-2 with SAT and VoxTell, 30 and 41 of 59 hypotheses are wrong in every repeat, often because gates refuse implausible readings.

Positive controls are easier than refutations: ACDC has 2% error on positive controls versus 19–20% on no-effect and hard-negative controls; UCSF-PDGM with VoxTell has 23% versus 48–61%. Inspecting wrong runs showed four recurring causes. On the eight UCSF-PDGM necrosis items, inaccurate measurements give the right direction but a much smaller effect than the expert masks, so the size test yields UNDERPOWERED or REFUTED where the reference is SUPPORTED. Cohort restrictions go wrong: a filter the hypothesis did not ask for, a missing grade filter, or a filter that keeps only one of the compared groups and yields no statistics. Planners name a predictor as a bare metric, such as `cdr_mean`, without the structure and observation the framework needs. Forms place the right variables in the wrong slots, such as a categorical association of sex with itself. These examples illustrate observed patterns; they are not an exhaustive partition. Missing statistical results are the largest INVALID category in every corpus, with similar rates across layer conditions.

G ADDITIONAL LIMITATIONS

Reference decisions. Some effect-size thresholds are conventions, and expert contours carry annotation error. PAPILA uses grader disagreement both as measurement information and to anchor δ; this creates a dependence between reference construction and uncertainty assessment. Replacing the graders' mean with one grader changes one of 67 patient-level reference labels and no eye-level label. The UCSF-PDGM comparison uses contrast-enhanced T1 images only; mask behavior varies by sequence. SAT's Dice requires reconstruction onto the expert grid, while its volumes are measured on its own grid.

Scope of the counterfactual. Paired and equivalence families, comprising thirty items, were not included in the injected-error study because the injection did not implement shifts for paired designs. Their counterfactuals therefore have no measured operating characteristics here. Changing the search target or the grid would create a new rule that needs its own evaluation.

H OFFLINE STUDIES OF THE MEASUREMENT LAYER

We tested shared instrument error, replacement error scales, and alternative checks after freezing the design corpus. None changed the deployed rule.

H.1 CORRELATED INSTRUMENT ERRORS

Instrument errors covary, especially for SAT and VoxTell. We compared instrument and expert volumes for 150 ACDC, 360 M&Ms-2, and 495 UCSF-PDGM subjects, by structure and cardiac timepoint, with Pearson correlations over 4,000 subject bootstraps. The fraction of summed error variance canceled in disagreement is $2\mathrm{Cov}(e_A, e_B)/(\mathrm{Var}(e_A) + \mathrm{Var}(e_B))$. Averaged over structures, this fraction is 0.80 for SAT and VoxTell on UCSF-PDGM and 0.49 on ACDC, but below 0.10 for every pair of instruments on M&Ms-2. It summarizes covariance without testing independence.

SAT and VoxTell have the most correlated errors on $11/14$ structure/timepoint combinations. On UCSF-PDGM, correlations are 0.87 [0.81, 0.90] for edema, 0.84 [0.76, 0.90] for whole tumor, 0.79 [0.60, 0.89] for enhancing tumor, and 0.70 [0.56, 0.83] for necrosis. Their shared fractions range from 0.70 to 0.86. BiomedParse generally shares less covariance with either tool.

Agreement between correlated instruments gives limited reassurance against shared error. Structure size can also correlate errors without a shared failure; this study does not separate the two.

H.2 AN EXPERT-DERIVED ERROR SCALE

Replacing instrument disagreement with expert-derived dispersion does not improve error ranking on UCSF-PDGM. We replaced it with the SD of instrument-minus-expert volume for each of the three instruments, holding the search and other checks fixed. This SD does not quantify systematic bias. Re-derivation first reproduced the stored results on all 1,552 eligible runs. Replacement applies to the 490 (31.6%) with a characterized biomarker; the others keep their original checks.

Relative to the expert-based SD, deployed scales are smaller by factors of 1.6–5.4 for VoxTell and 2.0–2.9 for SAT; BiomedParse ratios are 0.9–1.3. The replacement flags more runs with every instrument, but precision falls by one to three points; recall rises for two instruments and is unchanged for SAT. Pooled over the three instruments, the check flags 31.9% instead of 28.4% of runs, precision falls from 37.3 to 34.7%, recall rises from 35.5 to 37.2%, and AURC moves from 0.111 [0.066, 0.176] to 0.115 [0.068, 0.182]. Level 0 has no wrong runs under either scale.

A larger error scale crosses more decision boundaries without necessarily identifying which original conclusions are wrong. Measurement dispersion and conclusion-error ranking are therefore distinct objectives; these results establish neither scale as universally preferable.

H.3 ADDITIONAL CHECKS NOT ADOPTED

A flag for disagreement between the run's instrument and its secondary instrument that varies with the hypothesis's group or predictor can catch distortions that a constant offset would not. It closes 77% of the oracle gap on UCSF-PDGM with SAT but ranks worse than an uninformative score on four other pairs; as a fifth check it raises primary-instrument AURC from 0.031 to 0.037 [0.025, 0.050]. A voxel-level alarm, which flags Dice overlap with the three-instrument majority mask below the cohort median minus 0.1 median absolute deviation (Kofler et al., 2021), correlates with expert Dice on UCSF-PDGM (median point-biserial correlation -0.66) and flags all 19 scorable left–right-mirrored SAT masks on M&Ms-2. Over 3,146 eligible imaging runs, however, its AURC is 0.289 against 0.261 for an uninformative ordering and 0.210 for the deployed instrument check. Shared mask errors and missing common-grid comparisons limit such alarms.