MEDGAUGE: EVIDENCE-GROUNDED COMMIT PROTOCOLS FOR RELIABLE TOOL-USING LLM AGENTS IN MEDICAL MEASUREMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model agents increasingly operate the pipelines that turn medical images into reported measurements such as ventricular volume and cardiothoracic ratio. When such a pipeline errs, its output stays well formed and physiologically plausible, and an agent given quality signals still explains them away in favour of the value it holds. Runtime enforcement moves the decision to a program that blocks a violating action or substitutes a predefined one, but neither response states what evidence would make a blocked measurement reportable. We present MedGauge, a small commit layer that decides which values may be signed as trusted and which action would make a rejected value acceptable. Its observation channels compute evidence without executing the stage they audit, and sources that share an origin count as one witness. Each alarm names the call that clears it, which turns a blocked report into a repair the agent can execute, and a commit protocol accepts a report only when preconditions checked on the tool-call record hold. On a benchmark of 1,068 cases spanning cardiac MRI, abdominal CT and chest radiography, MedGauge signs no wrong value across the 812 covered fault-and-reasoner combinations of its two primary reasoners, and across four reasoners it releases none of 416 conflicts that the available evidence cannot resolve. It also achieves the highest usable output of any compared agent on every task (56% to 66%), because declared discharge conditions let the agent clear false alarms and repair faulty cases. On abdominal CT, repairs rise from 3 cases to 62. With channels, alarms and discharge conditions fixed, removing only the commit check produces 30 silent commits and 107 released conflicts. One adapter per task extends the layer from volumes to a two-dimensional ratio, and we release the code and the benchmark construction.

1 INTRODUCTION

Automated pipelines produce ventricular volume, ejection fraction, organ volume and cardiothoracic ratio, and these values enter diagnostic reports and trial endpoints directly. Language-model agents now run such pipelines end to end, from invoking the measurement tools to inspecting quality signals (Wang et al., 2026b; Fallahpour et al., 2025). We study the final step, the decision to commit a number, where the failure that matters leaves no trace. When the pipeline errs, its output is still well formed and physiologically plausible and the ratios derived from it stay self-consistent, so every downstream check passes: a silent failure (Bungert et al., 2023). The same problem arises whenever an agent reports a continuous quantity computed by its tools.

Figure 1 follows one cardiac MR study. On two basal slices the left ventricular mask extends into the left atrium, so the end-diastolic volume includes atrial blood. The value is still typical of a dilated ventricle, and the pipeline's checks pass (Figure 1a). An agent with the measurement tools remeasures the same mask and signs the value as trusted (Figure 1b). An agent that also has observation channels receives a shape alarm pointing to the leak, notes that it does not affect the result, and signs the same value (Figure 1c). The evidence was available, and the outcome depended on how the agent read it. Language models adopt the output of a silently failing tool and justify it (Sun et al., 2024).

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.

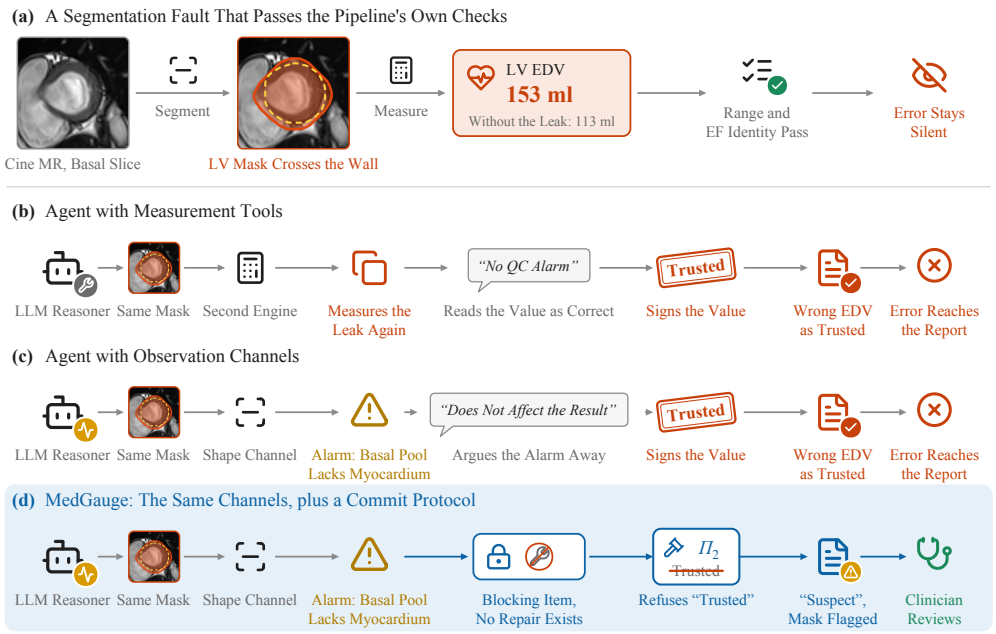


Figure 1: One silent fault in a cardiac MR study, followed through three verifiers. (a) The left ventricular mask crosses the ventricular wall (dashed) into the left atrium, and the pipeline's checks pass. (b) An agent with measurement tools re-measures the same mask and signs. (c) An agent that also has the observation channels explains the shape alarm away and signs. (d) In MedGauge the alarm opens a blocking item that declares no discharge, and the value is reported as suspect. Quoted notes paraphrase the recorded sessions.

Existing work answers this failure in two ways. The first gives the agent more tools or evidence, such as quality-control signals, verifiers or self-critique (Wang et al., 2026b; Fallahpour et al., 2025), but leaves the decision to the reasoner: with our full observation contract and every alarm, the two primary reasoners still sign 14 wrong cardiac values and release 104 of 160 conflicts that the evidence cannot resolve (Section 4.4). The second moves the decision into a program that blocks actions violating a policy or substitutes a predefined corrective action (Liu et al., 2026b; Wang et al., 2026a; Batole et al., 2026). Neither response states what evidence would make a blocked number reportable, and the two runtime-enforcement baselines in Table 1 repair at most 4% of injected faults.

Unlike a discrete action, a measurement is recoverable: a volume computed at the wrong calibration can be recomputed at the right one. What the agent lacks is an authoritative statement of which action recovers the number, based on evidence that holds however the reasoner argues. Such a statement can fail in three ways. The evidence may depend on the stage it is meant to judge, an alarm may report a problem without saying what would resolve it, and one record may be counted as several witnesses.

We therefore present MedGauge, a commit layer between the reasoner and the report that makes the commit decision: which numbers may be signed, and what evidence would allow a rejected number to be signed. It has three components, one for each failure.

Evidence comes from outside the stage it judges. A second measurement of a leaking mask reproduces the leak, because it reads the output it is meant to judge. Each channel therefore excludes the audited stage from its computation graph and declares, per family, the faults it covers.

An alarm names the evidence that clears it. An alarm that only reports a disagreement leaves the agent two options, releasing the value or downgrading the case. A blocking item that names the call that would clear it, with its parameters and tolerance, gives the agent an action to take.

Redundancy is counted by origin. With every channel available, as in Figure 1c, our two primary reasoners release 75 and 78 of the 80 unarbitrable geometry conflicts in the cardiac task. They argue

that three of four slice-thickness declarations agree, although those three are copies of one record (Appendix A). We therefore group geometry sources by provenance, with one witness per group.

A commit protocol ties the three components together. A program accepts a report only when three preconditions hold on the tool-call record and the structured fields of the report. The reasoner keeps every investigative choice, but arguing that an alarm is harmless cannot get a report accepted. In Figure 1d the shape alarm opens an item that declares no discharge, so the value is committed as suspect.

Our contributions are the following.

- **An observation contract that makes safety and usefulness compatible.** Each channel returns one of three states, and each alarm carries a declared scope and discharge condition. Added to an agent that already has the channels, the contract raises repairs on abdominal CT from 3 cases to 62 and usable output from 43% to 65% (Table 3).
- **A commit protocol that separates enforcement from information.** Its preconditions read the call record, and Proposition 1 classifies every route to a wrongly signed value. With the evidence fixed, removing the commit check yields 30 silent commits and 107 released conflicts.
- **An evaluation on 1,068 cases showing that, once a blocking item is raised, safety no longer depends on the reasoner.** It includes distractors, unarbitrable conflicts and fault families registered as uncovered. The two primary reasoners sign no wrong value across 812 covered fault-and-reasoner combinations, four reasoners release none of 416 unarbitrable conflicts, and MedGauge gives the highest usable output on every task. We release the code, including the fault injection that rebuilds both public-data tasks in full.

2 Related work

Five lines of work bear on who may sign a measured number (Appendix N).

Silent tool errors. Models accept a wrong number from a tool that raises no error and supply reasons for it (Sun et al., 2024; Ye et al., 2024). Remeasurement, post-hoc quality control and reference-free quality estimation return a new number or score that is then taken as the new truth (Heiman et al., 2025; Ruijsink et al., 2020; Cosarinsky et al., 2026; Marcianò et al., 2026). We run such estimators inside our observation layer as witnesses whose alarms must themselves be discharged.

Scaffold design and self-correction. Both leave the decision with the reasoner, whether they shape the loop, interface and tool set (Yao et al., 2023; Yang et al., 2024; 2025; Wang et al., 2026b; Fallahpour et al., 2025) or check the model's own output (Wang et al., 2023; Shinn et al., 2023; Madaan et al., 2023; Gou et al., 2024; Dhuliawala et al., 2024). Instructions remain the only control, and GPT-4 under a safety prompt still fails 23.9% of emulated test cases (Ruan et al., 2024).

Runtime enforcement. The closest line of work gates actions with a program, using Hoare contracts (Liu et al., 2026b), trigger-action policies (Wang et al., 2026a), instructions compiled to Prolog (Batole et al., 2026), least privilege (Shi et al., 2025), model-written policies (Xiang et al., 2025; Tsai & Bagdasarian, 2025), or rails, ontologies and isolation (Rebedea et al., 2023; Liu et al., 2026a; Wu et al., 2025). We agree that the gate must be a program, and these gates already express rich conditions: AgentSpec evaluates predicates and can invoke a parameterised corrective action, and ToolGate lets a tool result update the trusted state only when its post-condition holds. Our differences follow from the object being a measurement. Their conditions are written as a policy, an instruction or a contract over the state the agent exposes. Ours are computed by channels that exclude the stage they audit. They evaluate a continuous quantity together with its measurement parameters and provenance, and they hold only within a registered coverage. After a violation, our agent receives the call that clears it, with the tool, parameters and tolerance declared in the item. Section 4.4 approximates feedback without such a call by keeping every alarm and removing its discharge condition.

Measurement traceability. Traceability records where a number came from for a reader to act on (Daghyani et al., 2026; Long et al., 2026), and replayable imaging runtimes score the evidence an agent submits after the episode ends (Shen et al., 2026). Our blocking items veto the report before it is committed, until the named evidence clears them. We count silent errors as in the failure-detection literature (Bungert et al., 2023; Jaeger et al., 2023; Traub et al., 2024).

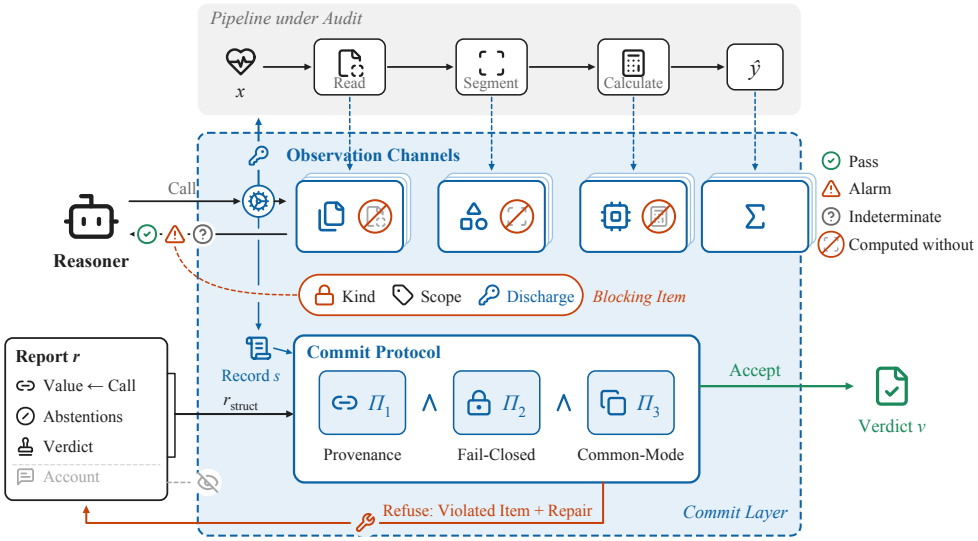


Figure 2: Overview of the MedGauge commit layer. The tinted region is the commit layer, the only part whose correctness the guarantees of Section 3.5 assume; the reasoner lies outside it. Each channel computes without the stage it audits (crossed icon). The layer executes every call and appends it to the record s. An alarm returns a blocking item, an obligation that locks reporting, and its discharge condition names the call that clears it (key). The commit protocol decides from s and the structured fields of the report (Equation 2). A refused report is returned with the violated item and an executable repair.

3 METHOD

3.1 OVERVIEW

MedGauge assigns the commit decision to a program and leaves the investigation to the reasoner. A case x enters the audited measurement pipeline $T = T_{\text{calc}} \circ T_{\text{seg}} \circ T_{\text{read}}$, which returns candidate values $\hat{y} = T(x)$. The reasoner then acts through two interfaces (Figure 2). It reads evidence by calling observation channels, each of which returns one of three states and any items that block reporting. It commits by submitting a report r, which names the call behind each reported field and gives the fields it abstains on, a verdict $\nu \in \{\text{trusted}, \text{suspect}, \text{unusable}\}$ and its own account of the case. To *commit* is to submit the report for adjudication, and a value is *signed* when it receives the verdict trusted; the reasoner commits and the program decides what is signed.

The reasoner runs a standard tool-calling loop (Yao et al., 2023) and makes every investigative choice. The program makes three decisions: which evidence exists (Section 3.2), what state that evidence is in (Section 3.3) and what may be signed (Section 3.4).

3.2 INDEPENDENCE AND COVERAGE

Evidence about a stage is informative only if it is computed without that stage. A second reading of a corrupted header reproduces the first exactly. Let g_c denote the computation of channel c, and let $\mathcal{J} = \{\text{read}, \text{seg}, \text{calc}, \text{out}\}$ contain the three stages of T and the reported output. Stage j is independent of channel c when T_j is absent from $\text{graph}(g_c)$, and the coverage of c is the subset of independent stages whose output the channel actually audits:

$$\text{indep}(c) = \{\, j \in \mathcal{J} \;:\; T_j \notin \text{graph}(g_c) \,\}, \qquad \text{cov}(c) \subseteq \text{indep}(c). \tag{1}$$

Independence makes evidence worth reading. Coverage is the stronger claim that the channel responds to errors of a stage, declared per fault family (Appendix E). A metadata check is independent of segmentation but covers none of it. Absence from the graph means that the channel does not execute

the stage. The channel may still read that stage's output as the object it audits, and coverage declares whether such a reading is informative. A second measurement of a leaking mask therefore has segmentation in indep but not in cov. Independence on the computation graph also leaves shared causes upstream. Two channels that both exclude T_{read} may read the same acquisition record, and versions written independently from one specification fail on the same inputs far more often than independence predicts (Knight & Leveson, 1986). The protocol therefore counts witnesses by provenance group (Section 3.4).

A useful channel set satisfies $\bigcup_c \mathrm{cov}(c) = \mathcal{J}$, so we place evidence sources at every stage: redundancy in the data at the input, anatomical priors and a second measurement implementation in between, and algebraic identities and physiological ranges at the output. A channel is adjudicating when its alarm rate on a fault family separates from its rate on natural cases. The protocol enforces only the states of adjudicating channels, and the other, advisory channels provide context the reasoner may cite.

3.3 Observation contract

Each channel returns pass, alarm or indeterminate. The third state is needed because a channel's own reference can fail, and a two-state contract would then force a choice between a false alarm and silence. It is returned only when the channel's applicability test fails, so a channel whose reference holds behaves exactly as under a two-state contract (Appendix E).

With its state, an alarm returns the items that block reporting. Each item is a triple of a kind, a scope and a discharge condition. The scope is what the item locks: the session, the reported handle or a named set of fields; the rest of the report stays available. The discharge condition names the evidence that clears the item: a tool, the parameters the call must carry, the comparison it must pass and a tolerance ε_b. Some kinds declare no discharge, such as a mask defect that no parameter can undo.

Declaring the condition distinguishes safety from abstention. A reasoner that sees only the alarm must choose between releasing the value and downgrading it, and our baselines do both. A reasoner that sees the condition knows which call clears the item, so a faulty case becomes repairable.

3.4 Commit protocol

A trusted verdict requires three preconditions, none of which reads the report's free text:

$$\mathsf{accept}(s, r) \iff \Pi_1(s, r) \wedge \Pi_2(s, r) \wedge \Pi_3(s, r), \qquad \Pi_i(s, r) = \Pi_i\big(s,\ r_{\mathrm{struct}}\big). \tag{2}$$

Here s is the layer's record of every executed call, and r_{struct} is the report without its free text: the reported fields with the calls they name, the abstentions and the verdict. Because each precondition depends only on these two objects, no explanation of why an alarm is harmless, however fluent, can change the decision. Appendix G maps the three preconditions onto eight implemented checks.

Provenance grounding (Π_1) requires each reported value to match, within 0.5%, the output of a successful call named in the report. That call must use the calibration defined for the report and carry no diagnostic-only parameter. These conditions prevent a value from certifying itself. Each task declares its diagnostic-only parameters, which are treated as a taint source with the submission as the sink (Newsome & Song, 2005). A value measured under one can be inspected but never signed.

Fail-closed on undischarged items (Π_2) refuses a trusted verdict while any item in scope of the reported fields remains undischarged. This is the classical fail-safe default (Saltzer & Schroeder, 1975), with each obligation arriving together with the evidence that clears it. An item is discharged by a call on the reported handle that carries the declared parameters and no diagnostic-only parameter, and that differs from the same call without those parameters by at most ε_b on every reported field (Appendix F). Three clauses close the routes around an alarm. Every adjudicating channel must have observed the reported handle. A channel that audits a value must have checked the value actually reported. A value reported from a handle other than the task's primary handle must be confirmed on the primary handle within ε_b; a second measurement path can thus confirm a value but cannot replace the value under audit. A repair passes all three, because the corrected handle is observed anew and items on the handle it replaces no longer apply once the report stops citing it.

Common-mode exclusion, Π_3, decides what counts as independent confirmation. Geometry sources are partitioned into witness groups by provenance, copies within a group count once, and the effective witness count is $n_{\text{eff}} = |\{k : G_k \neq \emptyset\}|$. When two groups disagree, arbitration requires a third group that agrees with one of them within the geometric tolerance, so $n_{\text{eff}} \geq 3$ is necessary. Otherwise the conflict is unarbitrable, and the admissible outcomes are a downgraded verdict with the values still reported or abstention on the affected fields (Appendix A).

A failed precondition returns the submission with the violated item and an executable repair, so the protocol limits what may be signed and leaves investigation free (Appendix H).

3.5 Guarantees and scope

Two guarantees follow from different premises. The first is mechanical: if the state record and the executor are correct, no trusted verdict is accepted while an item in scope of the reported fields remains undischarged. This holds by construction of Equation 2, and the undischarged-alarm release rate measures it directly (Section 4.2). The second concerns whether the item was raised at all.

Proposition 1. *Let a fault act on stage j with relative magnitude δ_ϕ on the reported fields. Suppose that some channel c with $j \in \mathrm{cov}(c)$ returns an alarm and declares a blocking item b on the handle used for the reported values, or on the primary handle those values substitute for. Suppose also that either b declares no discharge, or its tolerance satisfies $\varepsilon_b < \delta_\phi$ and its parameter test is sensitive to the fault, meaning an uncorrected fault of that magnitude separates the two compared calls by more than ε_b. Then $\nu = \mathrm{trusted}$ implies $\Pi_2 = 0$, so the submission is refused.*

Read in the contrapositive, the proposition lists the routes to a silent commit. A coverage hole leaves the affected stage without an alarming channel, and a specification hole produces an alarm that declares no item. The remaining routes are failures of tolerance, sensitivity or witnessing: $\varepsilon_b \geq \delta_\phi$, the declared parameter cancels out of the comparison, or copies within one group are counted as independent witnesses. A fault that scales every call equally, such as a global calibration error, is the sensitivity case, which is why such items declare no discharge. The tolerance, fixed at 1%, sets the trade-off between safety and usefulness explicitly. Appendix K locates each route in the evaluation.

The core of the layer is task-independent. A new task connects through six adapter hooks, two of which impose conditions on the pipeline: a second measurement implementation and a second geometry declaration (Appendix C). Eleven of the twelve registered channels serve several tasks, and chest radiography, a ratio on two-dimensional images, was added without a new channel class.

4 Experiments

4.1 Setup

We evaluate on three measurement tasks that differ in modality, dimensionality and the quantity reported. On cardiac MR the system reports left and right ventricular volumes, ejection fraction and myocardial mass over an institutional cohort of 504 cases. On abdominal CT it reports liver, spleen and kidney volumes over 276 cases from AMOS22 (Ji et al., 2022), auditing TotalSegmentator (Wasserthal et al., 2023). On chest radiography it reports the cardiothoracic ratio and its two widths over 288 cases built from the posteroanterior films of JSRT (Shiraishi et al., 2000) with the SCR annotations (van Ginneken et al., 2006). The audited pipeline is a keypoint graph model (Gaggion et al., 2023), with a pixel-level network (Cohen et al., 2022) as the second measurement path.

Fault injection. A retrospective cohort provides no reference for silent errors: a wrong volume in an archive is wrong by an unknown amount, at an unknown stage. We therefore inject faults into clean cases, which fixes the affected stage and magnitude, and use the output of the clean pipeline on the same case as ground truth. Mutation testing (Just et al., 2014), corruption benchmarks (Hendrycks & Dietterich, 2019) and evaluations of silent tool errors and segmentation quality control (Sun et al., 2024; Robinson et al., 2019) use the same protocol when real faults cannot be enumerated. Injection also builds two case types the evaluation needs. A distractor triggers the checks although the value remains correct. An unarbitrable conflict leaves two contradictory witnesses in the session, and the correct outcome is a downgraded verdict with the values reported. Cardiac MR holds 143 injected faults that move a reported value, 160 distractors, 80 conflicts and 120 natural cases; abdominal

Table 1: Safety and usefulness on three tasks with `qwen3.8-flash`, in percent. SE is over the injected faults of the covered families, and U over all cases except the unarbitrable conflicts and the boundary families (424, 256 and 268). Baselines are grouped by the families of Section 2 and receive our observation channels as plain alerts, without scope or discharge conditions; ReAct has only the task's measurement tools. Table A10 adds the boundary families, UAR and cost.

Method	Card. SE	FA	CR	Rep	U	CT SE	FA	CR	Rep	U	CXR SE	FA	CR	Rep	U
Scaffold design															
ReAct	31	18	82	2	56.8	28	4	100	3	43.0	27	14	42	0	42.5
MedAgent-Pro	6	39	85	0	46.2	7	4	0	0	41.0	6	14	0	0	38.1
MedRAX	5	47	74	1	42.0	0	12	0	0	39.5	3	19	25	1	37.7
Self-correction															
Self-consistency (k=5)	3	60	51	1	27.4	0	0	17	0	43.9	0	61	8	0	12.3
LLM-as-judge	3	68	33	2	28.1	0	70	0	1	20.4	2	25	0	0	33.6
Runtime enforcement															
GuardAgent	5	29	65	4	50.2	0	4	17	2	42.4	2	22	33	1	36.6
Progent	7	44	93	1	44.1	2	0	42	2	43.8	2	22	50	0	37.7
MedGauge (ours)	**0**	**0**	**0**	**8**	**64.6**	**0**	**0**	**0**	**46**	**65.2**	**0**	11	**0**	**37**	**62.3**

CT holds 131, 24, 12 and 100, and chest radiography 132, 36, 12 and 100. Each fault catalogue is based on documented failure modes of the audited pipeline (Appendix E). Abdominal CT and chest radiography each add a boundary family of 8 cases, registered as covered by no channel. Every method calls the same measurement tools, so the comparison isolates what happens between the tools and the report.

Metrics. Six case-level rates cover safety and usefulness (Appendix D). The silent error rate SE is the share of injected faults signed as trusted with a reported field outside tolerance, the false alarm rate FA the share of distractors whose correct value was withheld or downgraded, and the conflict release rate CR the share of unarbitrable conflicts signed as trusted. The strict repair rate Rep counts injected faults restored within a tolerance about three times tighter than the task's and signed, the usable output rate U counts trusted cases with every primary field within tolerance, and the undischarged-alarm release rate UAR is the share of trusted verdicts carrying an adjudicating alarm that no admissible evidence cleared. A policy that signs nothing scores $\mathrm{SE} = 0$ with $\mathrm{U} = 0$, so we report both.

Baselines and reasoners. Seven baselines cover the three families of Section 2 that place an agent in the loop (Appendix L), and any agent holding our observation channels counts as an ablation. Variation is measured by substituting the reasoner: `qwen3.8-flash` and `deepseek-v4-flash-0731` are the primary reasoners, and `GPT-4o-mini` and `Llama-3.3-70B` probe the weak end of the range. Tolerances are fixed per task (Appendix D), and every table reports the same configuration, with the seven obligations of Appendix I.

4.2 Main results

With the reasoner, cases and tools fixed, MedGauge achieves the highest usable output on every task (Table 1): 64.6% against 56.8% for the runner-up on cardiac MR, 65.2% against 43.9% on abdominal CT and 62.3% against 42.5% on chest radiography, with no silent error and no released conflict. The margin comes from discharge. On cardiac MR it comes mostly from distractor alarms cleared by their declared test, which leaves no false alarm, and on abdominal CT and chest radiography from repair. Baselines that approach our safety columns do so by downgrading faulty cases, and none of them repairs more than 4% of the injected faults on any task, whereas our system restores 8%, 46% and 37% of them within the strict tolerance and signs them. On cardiac MR the baselines trade silent errors against false alarms, from 5% and 47% for MedRAX to 3% and 60% to 68% for the two self-correction methods; our system lies off this trade-off (Figure A2).

The channels also catch errors nobody injected. On abdominal CT, the four natural cases whose two segmentation paths disagree beyond the threshold err by 29% to 98% on the flagged organ against the AMOS22 annotation, and our system abstains on it under every reasoner. Against expert annotations of all 200 natural CT and chest cases, the second path sees 11 pipeline errors, and none of the 44 resulting reasoner-and-field pairs is signed. Errors shared by both paths lie outside coverage. On clean cases our system downgrades 14% to 16%, with values still reported (Appendix M).

The undischarged-alarm rate shows how a correct value was reached. Across the twelve task-and-reasoner configurations of our system (Table 2), every in-scope blocking item is discharged in all

Table 2: Reasoner substitution with the prompt, cases and tools fixed. Counts are cases: SE over 143, 131 and 132 injected faults, FA over 160, 24 and 36 distractors, CR over 80, 12 and 12 conflicts; U and UAR are rates. The lower panel is ReAct + channels, the second row of Table 3.

Configuration and reasoner	Card. SE	FA	CR	U %	UAR %	CT SE	FA	CR	U %	UAR %	CXR SE	FA	CR	U %	UAR %
MedGauge (ours)															
qwen3.8-flash	0	0	0	64.6	0.0	0	0	0	65.2	0.0	0	4	0	62.3	0.0
deepseek-v4-flash	0	1	0	64.2	0.0	0	0	0	66.0	0.0	0	4	0	56.3	0.0
Llama-3.3-70B	0	156	0	1.7	0.0	0	3	0	35.5	0.0	13	11	0	40.7	0.0
GPT-4o-mini	0	152	0	4.2	0.0	0	13	0	31.6	0.0	3	19	0	32.1	0.0
ReAct + channels															
qwen3.8-flash	9	34	75	55.2	23.8	0	1	3	42.6	16.1	4	5	6	39.2	6.0
deepseek-v4-flash	15	30	78	57.3	23.1	0	4	2	45.3	19.0	14	6	1	39.6	4.1
Llama-3.3-70B	63	60	74	43.9	6.1	47	13	2	37.9	6.0	96	12	11	46.3	0.0
GPT-4o-mini	28	94	26	32.1	14.7	17	11	1	30.9	11.0	27	3	1	40.3	1.4

1,662 trusted reports, while on cardiac MR and abdominal CT every baseline that receives the same alarms signs despite some of them, in 3.5% to 27.0% of its trusted verdicts (Table A10). Each zero silent-error count has a one-sided 95% upper bound of 2.1% to 2.3%, and paired tests against an agent with the same tools but no observation layer give $p < 0.001$ on every task. Against MedRAX and GuardAgent the cardiac silent-error difference is not significant after Holm correction ($p = 0.078$), and the differences lie in conflict release, repair and usable output (Appendix D). On three additional samples totalling 403 cases, disjoint from the main benchmark in patients or in fault pairings, our system signs none of 672 covered combinations and releases no conflict (Appendix M).

4.3 Reasoner substitution

Table 2 substitutes four models from four vendors as the reasoner and compares our system with ReAct + channels, an agent that holds every channel and receives its alarms as plain alerts.

Whenever a blocking item is raised, safety is determined by the contract and the protocol. With both, the conflict column is zero in all twelve task-and-reasoner cells, 416 unarbitrable conflicts in total (104 per reasoner), and the silent-error column is zero in ten. Without them, the same reasoners produce 0 to 96 silent errors and release 1 to 78 conflicts. The two nonzero cells are the weaker reasoners on chest radiography. Of their 16 silent errors, 12 fall in one fault family, a lateral crop of the field of view, whose only responding evidence is an advisory hint that blocks nothing. There the outcome depends on the reasoner, and the criterion of Appendix I names this family as the next obligation to move into the program. The other four are Llama-3.3-70B errors: three heart leaks on which it skipped the second path, and a photometric inversion whose alarm declares no item (Appendix K).

The reasoner matters for usefulness. With identical blocking items and discharge conditions, strict repairs range from 0 to 62 cases per task, so turning a refusal into a repair is the reasoner's work, and the contract specifies the route. The two weaker models reach 31.6% and 35.5% on abdominal CT but fall to 4.2% and 1.7% on cardiac MR. There a volume can be computed under three calibration conventions, and the reasoner must identify the correct one before it can check the value; the weaker models read the resulting alarm as a discrepancy and downgrade. Without the contract and the protocol, the same models reach 32.1% and 43.9% there, at the cost of 28 and 63 silent errors. Usable output depends on the reasoner and the task; safety depends on the contract and the protocol.

4.4 Ablations

Every baseline so far has less information than our system. Table 3 removes this confound by adding the components one at a time, and Table 4 by removing each mechanism from the full system.

In Table 3, the channels remove silent errors by downgrading, and the contract turns those downgrades into repairs. The commit protocol uses the same evidence and alarms and adds only a program that checks the commit. It reduces undischarged-alarm release to zero with 31% lower cost and 35% fewer turns (Appendix H). Over the six configurations that share the information of the full system and lack only the commit check (Table A8b), removing the check produces 30 silent commits, 107 released conflicts and undischarged-alarm rates of 0.7% to 19.1%, against zero for the full system. The gap is largest on cardiac MR, where the contract alone releases 40 and 64 of the 80 unarbitrable conflicts.

Table 3: Cumulative ablation on abdominal CT with the same cases and the same reasoner, starting from a ReAct agent with the task's measurement tools. The second row is ReAct + channels of Table 2.

Configuration	SE %	CR %	Rep (cases / 131)	U %	UAR %
ReAct with the task's measurement tools	28	100	4	43.0	n/a
+ channels (§3.2)	0	25	3	42.6	16.1
+ observation contract (§3.3)	0	0	62	64.8	7.6
+ commit protocol (§3.4, **ours**)	0	0	60	65.2	**0.0**

Table 4: Leave-one-out ablation. Each row removes one mechanism from the full system and reports the column that the removal affects. Table A8 in Appendix J gives every column on every task.

Component	Mechanism removed	Task	Column	Full system	Removed
Evidence (§3.2)	Independence of witnesses	CT / CXR	UAR % (CT) / SE (CXR)	0.0 / 0	**5.4 / 1**
Contract (§3.3)	Declared discharge conditions	Abdominal CT	Rep (cases / 131) / U %	60 / 65.2	**28 / 52.7**
Protocol (§3.4)	Calibration conformance (Π_1)	Cardiac MR	SE (cases / 143)	0	**6**
Protocol (§3.4)	Provenance grounding (Π_1)	Cardiac MR	SE (cases / 143)	0	**7**
Protocol (§3.4)	Fail-closed on undischarged (Π_2)	Abdominal CT	UAR %	0.0	**5.4**
Protocol (§3.4)	Reporting-handle checks (Π_2)	Chest radiography	SE (cases / 132)	0	**6**
Protocol (§3.4)	Witness grouping (Π_3)	Cardiac MR	CR (cases / 80)	0	**37**

Each removal in Table 4 produces a different failure. Admitting a non-independent witness lets an auxiliary reading clear a blocking item. Removing the declared discharge conditions preserves safety but costs 32 repairs. Within the protocol, each precondition blocks a different route: without the calibration check of Π_1, six cardiac faults that scale every call equally are signed silently, and without the reporting-handle checks, six chest values from the second path are signed unconfirmed.

5 Discussion

Authority over the evidence. Interface design is an established factor for agents on software and web tasks (Yang et al., 2024; 2025). Our results isolate a second factor: who has authority over the evidence that crosses the interface. Every baseline except ReAct had the same channels, and in all 80 cardiac conflicts the inconsistency lay in the tool outputs, yet verdicts differed by scaffold.

A division of labour. The program decides what counts as evidence and when an item is cleared, and the reasoner decides how to obtain it, so a repair needs both a declared discharge condition and a model able to meet it (Section 4.3). The same division settles what to enforce. A rule written into a prompt holds only if the reasoner follows it. Once the program declares it, it becomes a property of the system. An obligation that one reasoner honours and another explains away therefore belongs in the program. Seven are enforced there on this basis (Appendix I), and the chest lateral crop is next (Appendix K).

Scope and next steps. Proposition 1 applies to faults covered by an adjudicating channel; outside coverage, the outcome depends on the reasoner. On the boundary families our system signs 1 of 8 CT cases and 6 of 8 chest cases (Table A10), and natural CT errors shared by both segmentation paths raise no alarm (Appendix M). A second path from a different model family would bring such errors into coverage. A channel that reports the calibration convention would move into the program the inference that weaker reasoners miss on cardiac MR. Resolving an unarbitrable conflict requires evidence from outside the session, such as the acquisition record.

6 Conclusion

We presented MedGauge, a commit layer that decides which values a tool-using agent may sign when it measures medical images. On 1,068 cases from three imaging tasks, the two primary reasoners sign no wrong value across 812 covered combinations, four reasoners release none of 416 unarbitrable conflicts, and MedGauge achieves the highest usable output of any compared agent, because declared discharge conditions let it clear false alarms and repair faults. The reliability of an agent that reports numbers is decided where values are committed, and a program can make that decision.

AI USE STATEMENT

We used generative AI tools for the subtasks below; for each, we state how its output was verified.

First, synthetic evaluation data. The fault catalogues of Appendix E are produced by deterministic injection code. An LLM assistant helped draft this code and proposed candidate fault families from a written taxonomy of measurement failures. The authors reviewed each candidate family and verified numerically, against the clean case, that its effect on the reported fields matched its specification; only families that passed this check are included in the benchmark.

Second, retrieval and discovery. We used an LLM assistant to search for related work and to identify candidate baselines. The authors checked every reference against the publisher record for its existence, author list and venue, and verified in the original paper each claim we attribute to it.

Third, writing. Drafts of the prose were produced, translated and polished with LLM assistance, and the authors rewrote, verified and approved every sentence.

We also used LLM assistance to implement parts of the system and the analysis scripts; this code is covered by unit tests and was reviewed by the authors.

We have reviewed all AI-assisted work described above and take responsibility for the final content of this paper.

ETHICS STATEMENT

The cardiac cohort consists of retrospective, de-identified clinical studies, used with the approval of the institutional review board of the institution that holds them; the committee name and approval number are withheld for anonymous review. The abdominal CT cases come from a public benchmark released under a CC BY 4.0 licence (Ji et al., 2022). The chest radiographs come from the JSRT database, used under its research-only terms, with annotations released under a CC BY 4.0 licence (Shiraishi et al., 2000; van Ginneken et al., 2006). No output reported here was used for a clinical decision, and a verdict of trusted describes the evidence in one session and is not a clinical endorsement.

REPRODUCIBILITY STATEMENT

The code of MedGauge, the three task adapters, the fault injection that builds the benchmark and every evaluated configuration is available in an anonymous repository at `https://anonymous.4open.science/r/medgauge-7142`.

REFERENCES

Algirdas Avizienis. The N-Version approach to fault-tolerant software. *IEEE Transactions on Software Engineering*, SE-11(12):1491–1501, 1985. doi: 10.1109/TSE.1985.231893.

Fraol Batole, Foutse Khomh, and Hridesh Rajan. RefineAct: Automatic runtime verification of LLM agent actions. In *Proceedings of the 41st IEEE/ACM International Conference on Automated Software Engineering (ASE)*, Munich, Germany, October 2026. ACM. To appear.

Till J. Bungert, Levin Kobelke, and Paul F. Jäger. Understanding silent failures in medical image classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14222 of *Lecture Notes in Computer Science*, pp. 400–410. Springer, 2023. doi: 10.1007/978-3-031-43898-1_39.

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P. Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. In *Proceedings of the 5th International Conference on Medical Imaging with Deep Learning (MIDL)*, volume 172 of *Proceedings of Machine Learning Research*, pp. 231–249. PMLR, 2022. URL `https://proceedings.mlr.press/v172/cohen22a.html`.

Matias Cosarinsky, Ramiro Billot, Lucas Mansilla, Gabriel Jimenez, Nicolás Gaggion, Guanghui Fu, Tom Tirer, and Enzo Ferrante. ConfIC-RCA: Statistically grounded efficient estimation of segmentation quality. *IEEE Transactions on Medical Imaging*, 45(7):3545–3557, 2026. doi: 10.1109/TMI.2026.3679618.

Matin Daghyani, Lyuyang Wang, Nima Hashemi, Bassant Medhat, Baraa Abdelsamad, Eros Rojas Velez, XiaoXiao Li, Michael Y. C. Tsang, Christina Luong, Purang Abolmaesumi, and Teresa S. M. Tsang. EchoAgent: Guideline-centric reasoning agent for echocardiography measurement and interpretation. *International Journal of Computer Assisted Radiology and Surgery*, 21(6): 1293–1300, 2026. doi: 10.1007/s11548-026-03694-0.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.212.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. MedRAX: Medical reasoning agent for chest X-ray. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15661–15676. PMLR, 2025.

Nicolás Gaggion, Lucas Mansilla, Candelaria Mosquera, Diego H. Milone, and Enzo Ferrante. Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: Applications to chest X-ray analysis. *IEEE Transactions on Medical Imaging*, 42(2):546–556, 2023. doi: 10.1109/TMI.2022.3224660.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Alice Heiman, Xiaoman Zhang, Emma Chen, Sung Eun Kim, and Pranav Rajpurkar. FactCheXcker: Mitigating measurement hallucinations in chest X-ray report generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30787–30796, 2025. doi: 10.1109/CVPR52734.2025.02867.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Paul F. Jaeger, Carsten T. Lüth, Lukas Klein, and Till J. Bungert. A call to reflect on evaluation practices for failure detection in image classification. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=YnkGMIh0gvX.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, and Ping Luo. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Advances in Neural Information Processing Systems 35 (NeurIPS), Datasets and Benchmarks Track*, volume 35, pp. 36722–36732, 2022. doi: 10.52202/068431-2661.

René Just, Darioush Jalali, Laura Inozemtseva, Michael D. Ernst, Reid Holmes, and Gordon Fraser. Are mutants a valid substitute for real faults in software testing? In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE)*, pp. 654–665. ACM, 2014. doi: 10.1145/2635868.2635929.

John C. Knight and Nancy G. Leveson. An experimental evaluation of the assumption of independence in multiversion programming. *IEEE Transactions on Software Engineering*, SE-12(1):96–109, 1986. doi: 10.1109/TSE.1986.6312924.

Zhuohang Li, Chao Yan, Nicholas J. Jackson, Wendi Cui, Bo Li, Jiaxin Zhang, and Bradley A. Malin. Towards statistical factuality guarantee for large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 11435–11456. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.576. URL https://aclanthology.org/2025.emnlp-main.576/.

Pengxiang Liu, Tao Ren, Wei Xiong, Tingrui Yang, Junjie Wang, and Jun Hu. OntoGuard: Enforcing action admissibility for LLM agents in complex interactive environments. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20937–20952. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.findings-acl.1051.

Yanming Liu, Xinyue Peng, Jiannan Cao, Xinyi Wang, Songhang Deng, Jintao Chen, Jianwei Yin, and Xuhong Zhang. ToolGate: Contract-grounded and verified tool execution for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9653–9684. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.findings-acl.470.

Ziyang Long, Xinqi Li, Junzhou Chen, Yifan Gao, Debiao Li, and Hsin-Jung Yang. BCER agent: Reliable long-horizon MRI workflow execution via compilation, artifact binding, and bounded local recovery. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, volume 16879 of *Lecture Notes in Computer Science*. Springer Nature Switzerland, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019.

Vincenzo Marcianò, Hava Chaptoukaev, Virginia Fernandez, M. Jorge Cardoso, Sébastien Ourselin, Michela Antonelli, and Maria A. Zuluaga. Diffusion-based quality control of medical image segmentations across organs. *IEEE Transactions on Medical Imaging*, 2026. doi: 10.1109/TMI.2026.3714697. Early access.

James Newsome and Dawn Song. Dynamic taint analysis for automatic detection, analysis, and signature generation of exploits on commodity software. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*. The Internet Society, 2005.

Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. NeMo Guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 431–445. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-demo.40.

Robert Robinson, Vanya V. Valindria, Wenjia Bai, Ozan Oktay, Bernhard Kainz, Hideaki Suzuki, Mihir M. Sanghvi, Nay Aung, José Miguel Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, Aaron M. Lee, Valentina Carapella, Young Jin Kim, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Chris Page, Paul M. Matthews, Daniel Rueckert, and Ben Glocker. Automated quality control in image segmentation: Application to the UK Biobank cardiovascular magnetic resonance imaging study. *Journal of Cardiovascular Magnetic Resonance*, 21:18, 2019. doi: 10.1186/s12968-019-0523-x.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL `https://openreview.net/forum?id=GEcwtMk1uA`.

Bram Ruijsink, Esther Puyol-Antón, Ilkay Oksuz, Matthew Sinclair, Wenjia Bai, Julia A. Schnabel, Reza Razavi, and Andrew P. King. Fully automated, quality-controlled cardiac analysis from CMR: Validation and large-scale application to characterize cardiac function. *JACC: Cardiovascular Imaging*, 13(3):684–695, 2020. doi: 10.1016/j.jcmg.2019.05.030.

Jerome H. Saltzer and Michael D. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975. doi: 10.1109/PROC.1975.9939.

Weixiang Shen, Chengzhi Shen, Yanzhu Hu, Che Liu, Junde Wu, Jiayuan Zhu, Xiao Han, Zongyue Li, Jingpei Wu, Min Xu, Daguang Xu, Yueming Jin, Benedikt Wiestler, Daniel Rueckert, and Jiazhen Pan. MedOpenClaw and MedFlowBench: Auditing medical agents in full-study workflows, 2026. arXiv:2603.24649.

Tianneng Shi, Jingxuan He, Zhun Wang, Hongwei Li, Linyu Wu, Wenbo Guo, and Dawn Song. Progent: Securing AI agents with privilege control, 2025. arXiv:2504.11703.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377.

Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000. doi: 10.2214/ajr.174.1.1740071.

Jimin Sun, So Yeon Min, Yingshan Chang, and Yonatan Bisk. Tools fail: Detecting silent errors in faulty tools. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14272–14289. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.790.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, volume 37, pp. 2323–2347, 2024. doi: 10.52202/079017-0076.

Lillian Tsai and Eugene Bagdasarian. Contextual agent security: A policy for every purpose. In *Proceedings of the 2025 Workshop on Hot Topics in Operating Systems (HotOS)*, pp. 8–17. ACM, 2025. doi: 10.1145/3713082.3730378.

Bram van Ginneken, Mikkel B. Stegmann, and Marco Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database. *Medical Image Analysis*, 10(1):19–40, 2006. doi: 10.1016/j.media.2005.02.002.

Haoyu Wang, Christopher M. Poskitt, and Jun Sun. AgentSpec: Customizable runtime enforcement for safe and reliable LLM agents. In *Proceedings of the 48th IEEE/ACM International Conference on Software Engineering (ICSE)*, pp. 2938–2950. ACM, 2026a. doi: 10.1145/3744916.3764546.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. MedAgent-Pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026b. URL https://openreview.net/forum?id=ZOuU0udyA4.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023. doi: 10.1148/ryai.230024.

Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. IsolateGPT: An execution isolation architecture for LLM-based agentic systems. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*. The Internet Society, 2025. doi: 10.14722/ndss.2025.241131.

Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Carl Yang, Dawn Song, and Bo Li. GuardAgent: Safeguard LLM agents via knowledge-enabled reasoning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68316–68342. PMLR, 2025.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, volume 37, pp. 50528–50652, 2024. doi: 10.52202/079017-1601.

Ke Yang, Yao Liu, Sapana Chaudhary, Rasool Fakoor, Pratik Chaudhari, George Karypis, and Huzefa Rangwala. AgentOccam: A simple yet strong baseline for LLM-based web agents. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=oWdzUpOlkX.

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

Junjie Ye, Sixian Li, Guanyu Li, Caishuang Huang, Songyang Gao, Yilong Wu, Qi Zhang, Tao Gui, and Xuanjing Huang. ToolSword: Unveiling safety issues of large language models in tool learning across three stages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2181–2211. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.119.

Serena Zhang, Sraavya Sambara, Oishi Banerjee, Julián Nicolás Acosta, L. John Fahrner, and Pranav Rajpurkar. RadFlag: A black-box hallucination detection method for medical vision language models. In *Proceedings of the 4th Machine Learning for Health Symposium (ML4H)*, volume 259 of *Proceedings of Machine Learning Research*, pp. 1087–1103. PMLR, 2025. URL https://proceedings.mlr.press/v259/zhang25c.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

A THE MOTIVATING CASE

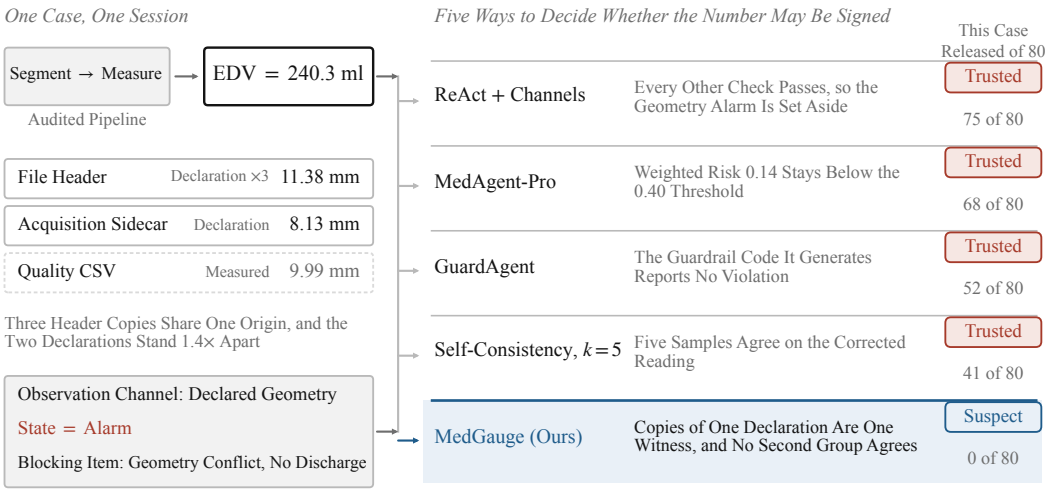


Figure A1: One case, one body of evidence, five decision procedures. Three copies of one record declare a slice thickness of 11.38 mm, the acquisition sidecar declares 8.13 mm, and a quality record measures 9.99 mm, which does not decide between them. Four scaffolds sign the volume as trusted, each with its own argument. The counts on the right are releases over all 80 unarbitrable conflicts of the cardiac task. Our protocol counts copies of one origin once, finds no second witness group that agrees, and returns a suspect verdict with the values still reported.

The case in Figure A1 is one of the 80 unarbitrable conflicts of the cardiac task and is the example Section 1 cites for witness counting. The geometry channel groups the four declarations by provenance, finds two witness groups, and raises an alarm. Its blocking item declares no discharge

condition, and this is accurate, because no measurement parameter decides which of the two records is stale. The releases on the right of Figure A1 are counted over all 80 cases with `qwen3.8-flash` as the reasoner, and the argument beside each row is condensed from the session logs. The same construction appears on the other two tasks as a contradiction between the sidecar and header spacing, with 12 cases each on abdominal CT and chest radiography.

B NOTATION

Three conventions in Table A1 deserve explicit mention. A verdict is committed on the submission as a whole, whereas the scope of a blocking item may be a single field. A session can therefore end with a trusted verdict and an abstained field. A channel state is a property of one call on one handle, so the state a report inherits is fixed by the session record. Finally, each task declares one primary handle, the measurement path that the task defines as the object under audit; the third clause of Π_2 refers to this handle.

Table A1: Notation used in Section 3 and Proposition 1.

Symbol	Meaning
s	session: the ordered record of tool calls and returns
r	submission: reported fields, a verdict, and the prose that argues for it
$\nu(r)$	verdict committed by the agent: trusted, suspect or unusable; the preconditions bind only $\nu =$ trusted
$\omega_i(s)$	state returned by channel i: pass, indeterminate or alarm
$B(s)$	blocking items in force, each carrying a scope and a discharge condition
Π_1, Π_2, Π_3	preconditions on the commit step, evaluated over s
ε_b	discharge tolerance of item b, 1% on the reported fields
h_{prim}	primary handle: the measurement path the task declares as the object under audit
$\mathrm{indep}(c)$	stages absent from the computation of channel c
$\mathrm{cov}(c)$	stages inside $\mathrm{indep}(c)$ whose output c audits and is sensitive to
n_{eff}	effective witness count: non-empty provenance groups among the geometry sources
SE, FA, U	silent error, false alarm and usable output rates
CR	conflict release rate: trusted verdicts over unarbitrable conflicts
UAR	undischarged-alarm release rate: trusted verdicts carrying an in-scope undischarged item

C TASK ADAPTERS AND THE SIX HOOKS

Section 3.5 states that the core of the commit layer is task independent and that a new task connects to it through a single adapter. This appendix lists the six hooks. Two of them are conditions on the pipeline, and a reader deciding whether their own setting qualifies should check these two first. The chest radiography column of Table A2 tests the claim on a task that differs from the other two in dimensionality and in the quantity reported.

For a pipeline that satisfies neither condition, the observation layer has empty coverage at the data and numeric layers. The protocol then reduces to provenance grounding, which still refuses a value that no call returned but cannot detect a wrong value that a call did return. The remaining four hooks make up the implementation cost of a new task. Eleven of the twelve registered channels serve more than one of the three tasks. Chest radiography reused the existing channel classes and needed three changes: a configurable spacing key on the geometry channel, a two-dimensional in-plane variant of the extent channel, and skipping the second formula where the quantity is a ratio of extents. The only single-task channel is the learned gross-error prior, fitted on a subset of the cardiac cases.

Table A2: The six adapter hooks. The two marked as pipeline conditions determine whether the observation layer has non-empty coverage; the other four are implementation work, and the table gives its size on our three tasks.

Hook	What it supplies	Kind	Cardiac MR	Abdominal CT	Chest radiography
Geometry declaration	a second record of pixel spacing and extent, of different origin from the header	**pipeline condition**	acquisition sidecar plus quality record	DICOM sidecar plus study index	NIfTI header, DICOM-style sidecar and two mask sidecars
Second measurement implementation	an independent path from mask to value, with a second formula	**pipeline condition**	slice summation against Simpson integration	voxel counting against surface integration	keypoint graph model against pixel-level CNN
Primary field set and tolerances	the fields a report must carry and the deviation each admits	implementation	4 fields, 15% relative and 5 points absolute	5 fields, 10% relative	3 fields, 10% relative
Anatomical prior	ranges and morphology the structure obeys	implementation	chamber volume ranges, wall continuity	organ volume ranges, tissue intensity	width ranges, lung-field contour continuity
Diagnostic-only parameters	the call arguments that must not reach a signed value	implementation	6 parameters	5 parameters	4 parameters
Fault catalogue	families, stages, magnitudes and coverage declarations	implementation	14 families	17 families	18 families

The second pipeline condition also carries a constraint that matters when the second path is heterogeneous. A second measurement implementation qualifies as a witness at any level of agreement with the primary path, but it qualifies as a source of reported values only within ε_b of the primary handle. On both volumetric tasks, the second path runs the same segmentation model at a second resolution and agrees with the primary path within 1% to 2%, so substituting one for the other changes little. On chest radiography, the second path is an independently trained network. At the median, its heart width is 5.4% narrower than that of the primary path, and its 95th-percentile disagreement is 16%. Removing the third clause of Π_2 allows 6 of 132 faults through (Table 4). The two roles of independence pull in opposite directions here. A second path from a different model family is the better witness, because its errors correlate less with those of the primary path. For the same reason, it is the worse replacement: the disagreement that makes it informative is a disagreement about the value itself. Declaring the second path strictly as a witness costs nothing when the paths agree and closes the substitution route when they disagree. It also lets coverage grow through witnesses from more distant model families, while the protocol restricts them to the witness role.

D METRICS AND STATISTICAL PROCEDURE

Write F_{prim} for the primary fields of a task and $F_{\mathrm{prim}}^{\mathrm{eval}}$ for the primary fields the task configuration leaves to the method, that is, F_{prim} without the fields every method abstains from by default. Write F_{rep} for the fields the submission reports, F_{abst} for the fields it abstains from, v_f for a reported value, $y_f^\star$ for the value the clean pipeline produces on the same case, and τ_f for the field tolerance, which is relative on volumes and widths and absolute on ejection fraction. Let ν be the committed verdict

and d_f the deviation in which the tolerance is stated, namely $|v_f - y_f^\star|/|y_f^\star|$ on a relative field and $|v_f - y_f^\star|$ on an absolute one. The metrics are then indicator functions on a case:

$$\mathrm{SE}(x) = \mathbf{1}\Big[\nu = \textsf{trusted} \ \wedge \ \exists f \in F_{\mathrm{rep}} : \ d_f > \tau_f\Big],$$

$$\mathrm{FA}(x) = \mathbf{1}\Big[\nu \neq \textsf{trusted} \ \vee \ F_{\mathrm{abst}} \cap F_{\mathrm{prim}}^{\mathrm{eval}} \neq \emptyset\Big], \qquad x \in \mathcal{X}_{\mathrm{distractor}},$$

$$\mathrm{U}(x) = \mathbf{1}\Big[\nu = \textsf{trusted} \ \wedge \ F_{\mathrm{prim}}^{\mathrm{eval}} \subseteq F_{\mathrm{rep}} \ \wedge \ \forall f \in F_{\mathrm{prim}}^{\mathrm{eval}} : \ d_f \leq \tau_f\Big],$$

$$\mathrm{CR}(x) = \mathbf{1}\big[\nu = \textsf{trusted}\big], \qquad x \in \mathcal{X}_{\mathrm{conflict}},$$

with $|\mathcal{X}_{\mathrm{conflict}}| = 80$ on the cardiac task and 12 on each of the other two tasks. The constant policy $\nu \equiv \textsf{suspect}$ satisfies $\mathrm{SE} \equiv 0$ and $\mathrm{U} \equiv 0$. Reporting SE alone therefore does not identify a useful system, which is why every table carries both metrics. The strict repair rate counts an injected fault as repaired when every affected primary field returns to within a tolerance about three times tighter than τ_f and the verdict is trusted.

The undischarged-alarm release rate is the fraction of trusted verdicts in which an adjudicating channel was in alarm on the reported handle and no legal evidence cleared the item. It separates a correct value reached through evidence from a correct value reached by overriding an alarm. It is also the metric that tests Π_2 directly, since a correct implementation of fail-closed makes it identically zero. Two definitional choices affect how the column should be read. First, the adjudicating set consists of `metadata_check`, `mask_stats` and `recompute_check`, the three channels that declare blocking items. An uncleared alarm on an advisory channel is not counted, because an advisory channel creates no obligation for the protocol to enforce. Second, the metric honours the scope of a blocking item. An item that names a set of fields binds only those fields, so a session that abstains from every field the item names has acted on the alarm and does not count as a release. Both choices follow the contract of Section 3.3. The run reports carry the per-session state of every channel, so the rate can be recomputed under either convention.

We compare paired binary outcomes on the same cases with the exact McNemar test, and release rates with Fisher's exact test. A test family is one metric on one task and contains every comparison we report between our system and another method. The cardiac silent-error family therefore holds eleven tests, the abdominal CT family eight, and the chest radiography family twelve. The p values in the text are Holm-corrected within each family, and the run reports carry the uncorrected values and the discordant counts from which the corrected values follow. Two comparisons change status under the correction. On cardiac silent errors, the comparisons against MedRAX and against the guardrail baseline each rest on seven discordant cases and give $p = 0.016$ uncorrected and $p = 0.078$ corrected; we report both as not significant.

For a zero count we report the one-sided 95% Clopper-Pearson upper bound: 2.1% on 143 cardiac faults, 2.3% on 131 covered CT faults, 2.2% on 132 chest faults, 0.74% on the 406 covered faults pooled across tasks for one reasoner, and 1.5% on the 202 injected replication cases. Pooled bounds are computed over cases, because the readings of one case by several reasoners are not independent.

The tables carry no error bars. We assess variation by substituting the reasoner (Section 4.3) and through the additional samples of Appendix M, and every zero count carries the exact upper bound given above. On the CT task, one case carries an anatomy-dependent injection that is infeasible on that case's anatomy, which leaves 275 of 276 sessions completed for every method. On chest radiography, two base studies were replaced in the manifest for the same reason.

Tolerances are set by each task: 15% relative on ventricular volumes and 5 points absolute on ejection fraction for the cardiac task, 10% relative on organ volumes for the CT task, and 10% relative on the cardiothoracic ratio and the two widths for the chest task, with the strict repair tolerances at 5%, 2 points and 3%. Every method, including the baselines, abstains from right ventricular fields by default, because the segmentation model under audit is unreliable on them. This decision belongs to the task configuration. These fields therefore lie outside $F_{\mathrm{prim}}^{\mathrm{eval}}$, and no method is charged a false alarm or denied usable output for abstaining from them.

E OBSERVATION CHANNELS AND FAULT FAMILIES

Table A3 lists the registered channels. Three of them are adjudicating: their state creates obligations that the protocol enforces. The rest are advisory and supply context that the reasoner may cite. Each channel declares the stage it audits and excludes that stage from its own computation. Each also declares an applicability domain and returns indeterminate when its own reference fails. Tables A4a to A4c give the three fault catalogues, together with the channels that declare coverage of each family. Faults that no adjudicating channel covers are still bounded by the preconditions of Section 3.4: a value that no measurement call returned fails Π_1 whatever the channels report, and a conflict with two witness groups fails Π_3.

Table A3: The twelve registered channels. The Layer column names the stage of the audited pipeline whose output the channel judges; each channel's computation excludes that stage. Adjudicating channels declare blocking items; advisory channels declare none.

Channel	Layer	Role	Evidence it computes	Tasks
`metadata_check`	data	adjudicating	geometry declarations from header, sidecar and quality record, grouped by provenance; a conflict between two witness groups blocks the session and declares no discharge	all three
`coverage_check`	data	advisory	measured extent against an anatomical reference; indeterminate when the reference view itself fails	all three
`mask_stats`	shape	adjudicating	morphology of the mask alone: interior gaps, dropped slices, leaks, detached components, label swaps, and cross-path disagreement	all three
`appearance_check`	shape	advisory	image intensity inside each structure against a tissue prior	CT, chest
`volume_curve`	shape	advisory	volume-against-frame diagnostics: single-frame spikes, adjacent extrema, collapsed amplitude	cardiac, CT
`recompute_check`	numeric	adjudicating	a second implementation with a second formula, compared against the value the report will carry	all three
`cross_check_summary`	numeric	advisory	spread across repeated measurements of one quantity under different parameters, with each value traced to the field it was returned under	all three
`provenance`	numeric	advisory	measurement ledger: per-slice contributions, contributing slice set, voxel volume, spacing	cardiac, CT
`identity_check`	output	advisory	algebraic identities among the reported fields, reading no image	all three
`plausibility`	output	advisory	physiological ranges and sentinel values, reading no image	all three
`percentile`	output	advisory	position of the value in the cohort distribution	all three
`risk_estimate`	output	advisory	learned prior combining weak signals into a gross-error probability	cardiac

We audit silence on the bare tool returns, with no agent involved, so that calling a family silent rests on measurement. For each of seven cardiac fault families, we fitted a logistic regression on the nine numeric fields of the bare tool returns, with no observation channel among the features, and scored it by stratified five-fold cross-validation against 45 natural cases. A family with an AUC at or below 0.6 was declared silent: interior gap at 0.31, double calibration at 0.42 and basal leak at 0.54. The other four families score above that threshold. We retain them because the agent holding only the measurement tools signs every case of all four, so their statistical separability is available in principle and unused in practice.

Table A4a: Cardiac MR fault families. The Channels column lists the channels that the fault specification registers for each family. Counts are cases in the case manifest; natural cases carry no family. Distractors leave the reported value correct, and conflicts leave two contradictory witnesses in the session.

Family	Injected error	n	Stage	Role	Channels
F1	Interior gap in the myocardial mask	14	shape	fault	`mask_stats`
F2	Basal or apical slices dropped	14	shape	fault	`coverage_check`, `mask_stats`
F3	Mask leaks into the atrium	14	shape	fault	`mask_stats`
F4	End-diastolic frame taken from the wrong phase	14	numeric	fault	`recompute_check`
F5	Calibration applied twice	14	numeric	fault	`recompute_check`
F6	Reported field scaled by ten	12	numeric	fault	`recompute_check`, `plausibility`
F10	Sidecar frame count disagrees with the header	12	data	fault	`metadata_check`
F13	Ejection fraction computed from the wrong formula	12	numeric	fault	`identity_check`, `recompute_check`
F14	Reported field divided by ten	12	numeric	fault	`recompute_check`, `plausibility`
F15	Left and right labels swapped	12	shape	fault	`mask_stats`
F17	End-systolic frame taken from the wrong phase	14	numeric	fault	`recompute_check`
F20	Reference range polluted, value correct	80	data	distractor	`coverage_check`
F22	Frame jitter, value correct	80	shape	distractor	`volume_curve`
F12	Slice spacing disagrees across sources	80	data	conflict	`metadata_check`

Totals: 144 injected faults (143 of which move a reported value), 160 distractors, 80 conflicts and 120 natural cases, 504 in all.

Totals: 132 injected faults (131 of which are feasible on their anatomy), 8 boundary cases outside coverage, 24 distractors, 12 conflicts and 100 natural cases, 276 in all.

Totals: 132 covered faults, 8 boundary cases, 36 distractors, 12 conflicts and 100 natural cases, 288 in all.

Two rows of Table A4c are worth reading against the design. K3 is the family whose only responding channel is advisory, so its silent count depends on the reasoner and ranges from 0 to 9 of 12 (Section 4.3). X3 swaps the two lung labels but leaves their union unchanged, so all three reported fields remain correct. The channel raises an alarm and declares no blocking item, because none of the three quantities depends on which lung is which.

F Discharge of a blocking item

A blocking item b carries a declared tool, the parameters that the discharging call must carry, and a tolerance ε_b. Let F_h denote the fields of the handle to which the item is scoped, and $\mathcal{K}_{\mathrm{diag}}$ the diagnostic-only parameters of the task. The item is discharged in session s when

Table A4b: Abdominal CT fault families.

Family	Injected error	n	Stage	Role	Channels
G1	Interior gap in the organ mask	12	shape	fault	`mask_stats`
G2	Superior slices dropped	12	shape	fault	`coverage_check`, `mask_stats`
G3	Spleen mask leaks into the stomach	12	shape	fault	`appearance_check`, `mask_stats`
G4	Left and right kidney swapped	10	shape	fault	`appearance_check`, `mask_stats`
G7a	Slice thickness overwritten in the header	12	data	fault	`metadata_check`
G11a	Volume returned in mm³ labelled ml	10	numeric	fault	`plausibility`, `recompute_check`
G13	Resampling scale inconsistent	12	numeric	fault	`recompute_check`
G15	Partial read of the volume	12	numeric	fault	`recompute_check`
G5	Spleen absent from the label map	8	shape	fault	`mask_stats`, `plausibility`
G6	One-voxel erosion of the left kidney	8	shape	**boundary**	*none registered; the cross-path rule of Table A7b responds*
G11b	Volume returned in litres labelled ml	8	numeric	fault	`plausibility`, `recompute_check`
G12	Label indices shifted	8	numeric	fault	`identity_check`, `plausibility`, `recompute_check`
G14	Total kidney volume from the wrong formula	8	numeric	fault	`identity_check`, `recompute_check`
G16	Border sentinel value counted as tissue	8	numeric	fault	`plausibility`, `recompute_check`
X2	Organ touches the field-of-view edge, value correct	12	data	distractor	`coverage_check`
X5	Detached accessory spleen, value correct	12	shape	distractor	`mask_stats`
X1	Sidecar spacing contradicts the header	12	data	conflict	`metadata_check`

$$\mathrm{disch}(b, s) = \mathbf{1}\Big[\exists\, c, c_0 \in s :\ \mathrm{tool}(c) = b.\mathrm{tool},\ \ b.\mathrm{args} \subseteq \mathrm{args}(c),\ \ \mathrm{args}(c) \cap \mathcal{K}_{\mathrm{diag}} = \emptyset,$$

$$\mathrm{tool}(c_0) = \mathrm{tool}(c),\ \ \mathrm{handle}(c_0) = \mathrm{handle}(c),\ \ \mathrm{args}(c_0) = \mathrm{args}(c) \setminus b.\mathrm{args},$$

$$\max_{f \in F_h} \frac{|v_f(c) - v_f(c_0)|}{\max(|v_f(c_0)|, \epsilon_0)} \le \varepsilon_b\Big],$$

where c_0 is the same call on the same handle with the declared parameters removed and every other argument unchanged, and ϵ_0 is a floor of one voxel volume that keeps the ratio defined when a structure is absent from the label map. The comparison is between the parameterised call and the plain one, which prevents a reasoner from reporting the parameterised value and then certifying that value against itself. The variant that compares against the reported value is the configuration marked *self-comparison admitted as discharge* in Table A8.

For the detached accessory spleen, for example, the declared discharge is a remeasurement with the largest-component parameter enabled, whose reported fields differ by at most 1% from a measurement without it. This establishes that the detached part is irrelevant to the reported quantity.

The equation above is an invariance test. It is the route by which a distractor alarm clears on the handle that raised it. Two other routes also clear an item, and the counts in Table 1 include all three. A genuine fault is repaired by supersession: the reasoner builds a corrected handle, every adjudicating channel observes that handle anew, and the report names it. No item of the superseded handle then

Table A4c: Chest radiography fault families: 25 specifications across 18 families. The family marked as having no adjudicating channel is the one that Sections 4.3 and 5 report as reasoner-dependent. It is also the only covered family, on any task, whose sole responding channel is advisory.

Family	Injected error	n	Stage	Role	Channels
K1	Heart mask leaks into the lung field	12	shape	fault	`mask_stats`
K2	Lung field cut laterally in the mask	12	shape	fault	`mask_stats`
K7	Heart shadow undersegmented, gap at the border	12	shape	fault	`mask_stats`
K3	Field of view cropped laterally in the image	12	data	fault	`coverage_check` *(advisory only)*
K4	Pixel spacing overwritten in the header	12	data	fault	`metadata_check`
K8	Photometric interpretation inverted	10	data	fault	`metadata_check`, `appearance_check`
K5	Width reported in centimetres labelled millimetres	10	numeric	fault	`recompute_check`
K6	Readout scaled by the wrong factor	12	numeric	fault	`recompute_check`
H1	Heart label absent from the map	8	shape	fault	`mask_stats`
H3	Ratio computed with numerator and denominator exchanged	8	numeric	fault	`identity_check`, `recompute_check`
H4	Stale cardiothoracic ratio carried over	8	numeric	fault	`recompute_check`
H5	Heart width replaced by a sentinel value	8	numeric	fault	`plausibility`, `recompute_check`
H6	Thorax width measured on one lung only	8	numeric	fault	`identity_check`, `recompute_check`
H2	Heart shadow dilated uniformly by 10 pixels	8	shape	**boundary**	*none*
X1	Sidecar spacing contradicts the header	12	data	conflict	`metadata_check`
X2	Field of view cropped inferiorly, widths correct	12	data	distractor	`coverage_check`
X3	Left and right lung labels swapped, union unchanged	12	shape	distractor	`mask_stats` *(alarm, no blocking item)*
X5	Detached satellite component beside the heart, value correct	12	shape	distractor	`mask_stats`

remains in force under Π_2. A repair therefore has to move the value, and the invariance test is never asked to accept it. A recomputation mismatch clears by the third route: the reported value agrees with an independent calculation path in the reported convention. This is the rule listed first in Table A7b.

Supersession and substitution look alike in the record, but they differ in what confirms the new value. A superseding handle is a corrected version of the primary handle, produced by the same measurement implementation under different parameters. The primary handle's own measurement therefore stays available as the confirmation that the third clause of Π_2 requires. For example, removing a 0.9 ml detached component from a 200 ml organ leaves the two measurements within 0.5%. A substituted handle comes from a different measurement implementation. Either the confirmation exists within ε_b, and the substitution then changes nothing that matters, or it does not. In the second case, the disagreement between the two paths is itself the finding, and the admissible outcomes are a downgraded verdict or abstention. The clause therefore permits every repair that the invariance test already permits, and it refuses only the route that carried no independent check on accuracy.

G FROM THREE PRECONDITIONS TO EIGHT CHECKS

The paper states three preconditions, grouped by the kind of unreportable value that each one excludes. The implementation runs eight checks, because a single precondition can be violated along several routes that need separate tests. Table A5 gives the correspondence.

Table A5: The three preconditions of Section 3.4 against the checks in the implementation. The last Π_2 row is the seventh obligation of Appendix I.

Precondition	Check	What it refuses
Π_1	provenance grounding	a reported value that matches no named successful call within 0.5%
Π_1	diagnostic-mode taint isolation	a value whose source call carried a diagnostic-only parameter
Π_1	calibration conformance	a value measured at a calibration other than the reported one
Π_2	fail-closed on blocking items	a trusted verdict while an item in scope is undischarged
Π_2	all-channels check	a reported handle that some adjudicating channel never observed
Π_2	value-level verification	a channel that checked a value other than the reported one
Π_2	primary-handle confirmation	a value from a substituted handle that the primary handle does not confirm within ε_b
Π_3	common-mode exclusion	copies of one record passing as independent witnesses

Each check closes one route by construction. How often a route is attempted determines how visible its check is in the ablations of Appendix J. Diagnostic-mode taint isolation fired twice across the volumetric tasks; otherwise the reasoners kept diagnostic calls out of their reports. Removing this check alone therefore leaves the cardiac columns of the full system unchanged. The all-channels check, value-level verification and primary-handle confirmation share one switch in the implementation, so the single-factor ablation removes all three together. Its effect concentrates on chest radiography, where the ablation returns 6 of 132 faults to silence through the substitution route that the third clause of Π_2 closes. Value-level verification also fired once on its own, on an additional CT sample (Table A7b).

H PER-REASONER COST, TURNS AND REFUSALS

Table A6 reports the budget columns per reasoner for our system and for ReAct + channels, which has the same channels but neither the contract nor the protocol. The table answers two questions. The first is whether a refused submission consumes budget. A refusal itself is free, and the cost column counts every model turn a session consumed, including the turns that followed a refusal, so the two configurations are compared on the same basis. The second is whether the protocol taxes investigation. With `qwen3.8-flash`, the protocol lowers both cost and turns on all three tasks, as it does with `deepseek-v4-flash` on abdominal CT, because a reasoner that knows which call clears an item stops arguing and makes that call. Llama-3.3-70B and GPT-4o-mini spend more under the protocol, and they also trigger most of the refusals.

With the primary reasoner, at least one submission is refused in 5.2%, 2.9% and 4.2% of sessions on the three tasks. The counts rise as the reasoner weakens and peak at 184 of 504 sessions for the weakest model in the set. This model reports values that it never put through the per-value check, or reports them in the uncorrected calibration convention. Every session reaches an accepted submission. The two guardrail baselines lie outside the scale of this table: GuardAgent costs 1.00, 0.62 and 0.34 CNY per case on the three tasks, 26, 18 and 11 times the cost of our system, because it generates and runs check code in each session.

On abdominal CT with `qwen3.8-flash`, the full system costs 0.035 CNY and takes 11.8 turns per case, against 0.051 CNY and 18.2 turns for the contract-only configuration. These are the 31% and 35% reductions reported in Section 4.4. On chest radiography, our system costs 0.032 CNY per case

Table A6: Budget and refusal behaviour. Refusals counts submissions returned because a precondition failed; Sessions refused counts sessions with at least one refusal. Downgraded natural counts the natural cases that receive a verdict other than trusted while their values are still reported. The chest run report records per-case cost only for a subset of configurations; a dash marks a cell that the report does not include.

Task	Configuration	Reasoner	Cost (CNY / case)	Turns / case	Refusals	Sessions refused	Downgraded natural
Cardiac	MedGauge	qwen3.8-flash	0.038	12.1	26	26 / 504	17 / 120
Cardiac	MedGauge	deepseek-v4-flash	0.059	7.2	14	14 / 504	15 / 120
Cardiac	MedGauge	Llama-3.3-70B	0.132	16.0	219	184 / 504	117 / 120
Cardiac	MedGauge	GPT-4o-mini	0.048	9.7	99	94 / 504	110 / 120
Cardiac	ReAct + channels	qwen3.8-flash	0.045	17.3	n/a	n/a	19 / 120
Cardiac	ReAct + channels	deepseek-v4-flash	0.045	7.5	n/a	n/a	18 / 120
Cardiac	ReAct + channels	Llama-3.3-70B	0.067	10.5	n/a	n/a	30 / 120
Cardiac	ReAct + channels	GPT-4o-mini	0.024	7.1	n/a	n/a	50 / 120
CT	MedGauge	qwen3.8-flash	0.035	11.8	9	8 / 276	16 / 100
CT	MedGauge	deepseek-v4-flash	0.044	6.2	7	7 / 276	15 / 100
CT	MedGauge	Llama-3.3-70B	0.117	13.0	120	86 / 276	33 / 100
CT	MedGauge	GPT-4o-mini	0.050	8.6	30	19 / 276	30 / 100
CT	Contract only, no protocol	qwen3.8-flash	0.051	18.2	n/a	n/a	18 / 100
CT	ReAct + channels	qwen3.8-flash	0.052	17.9	n/a	n/a	16 / 100
CT	ReAct + channels	deepseek-v4-flash	0.051	7.6	n/a	n/a	10 / 100
CT	ReAct + channels	Llama-3.3-70B	0.049	7.5	n/a	n/a	18 / 100
CT	ReAct + channels	GPT-4o-mini	0.035	8.3	n/a	n/a	33 / 100
Chest	MedGauge	qwen3.8-flash	0.032	13.6	12	12 / 288	14 / 100
Chest	MedGauge	deepseek-v4-flash	—	7.1	39	—	27 / 100
Chest	MedGauge	Llama-3.3-70B	—	14.6	86	—	17 / 100
Chest	MedGauge	GPT-4o-mini	—	8.2	66	—	31 / 100
Chest	Contract only, no protocol	qwen3.8-flash	0.041	18.9	n/a	n/a	28 / 100
Chest	ReAct + channels	qwen3.8-flash	0.041	19.0	n/a	n/a	27 / 100
Chest	ReAct + channels	deepseek-v4-flash	—	9.5	n/a	n/a	—
Chest	ReAct + channels	Llama-3.3-70B	—	6.0	n/a	n/a	—
Chest	ReAct + channels	GPT-4o-mini	—	7.9	n/a	n/a	—

and takes 13.6 turns, against 0.041 CNY and 19.0 turns for ReAct + channels. The same ordering holds on abdominal CT. On cardiac MR it reverses for one of the two reasoners, because the repair route that the protocol opens is used more often there and each repair costs turns.

I THE ENFORCEMENT CRITERION AND THE SEVEN OBLIGATIONS

We first state the criterion for moving an obligation out of the prompt, since it determines which obligations this section lists. A rule written into the prompt is a joint property of the rule and the reasoner; once the program declares it, the rule becomes a property of the system. Table A7a measures this difference on one obligation, and the procedure follows from it: run a second reasoner under the same prompt, and the obligations that break are those that belong in the program.

Table A7a: The same obligation on cardiac MR, either stated in the prompt as an advisory alert or declared by the program as a blocking item, with the rest of the prompt fixed. Silent errors over 143 injected faults.

Reasoner	Stated in the prompt	Declared by the program
qwen3.8-flash	0	0
deepseek-v4-flash	5	0

When the recomputation mismatch is demoted from a blocking item to a prompt-level obligation, qwen loses nothing but deepseek commits 5 silent errors; with the blocking item declared, neither commits any. One single-factor configuration makes the point most clearly. Removing only the witness-group blocking item raises cardiac conflict release from 0 of 80 to 37 of 80 ($p < 0.001$), even though the prompt still contains the sentence that forbids copies of one record from counting as a majority.

The criterion also identifies the next obligation. On chest radiography, the lateral field-of-view family has no adjudicating channel, and silent commits are 0, 0, 3 and 9 of 12 across the four reasoners, against 10 of 12 for the unprotected ReAct baseline. This spread is the pattern the criterion looks for, so the family points to the next obligation to move into the program. Table 2 reports its cost, and its route is the first situation of Proposition 1.

Table A7b lists the seven rules declared on the basis of evidence of this kind, each with the failure it targets and the outcome with the rule in force. All thresholds in these rules are calibrated on clean cases. The development reasoner named in the table is a fifth model that we used while building the system; it appears in no comparison table.

We apply the rules that act at the call boundary or at the commit check to every recorded session of the strict protocol. Reported sessions whose outcome a rule changes are executed with the rule in force. Every table therefore reports the configuration that carries all seven rules.

Table A7b: Seven obligations declared by the program. Each targets the same pattern: an obligation stated as an instruction holds on one reasoner and gives way on another, or on the same reasoner when it follows a different line of argument. The middle column gives the failure without the rule, and the right column gives the outcome with it. On clean cases, every rule costs zero false alarms except the third, which abstains on the flagged organ in 4 of 100 natural CT cases; each of these cases carries an error of 29% to 98% on that organ against the AMOS22 expert annotation.

Declared rule	Failure without the rule	With the rule
A recomputation mismatch blocks the report, discharged by re-checking the reported value in the reported convention	deepseek 5 silent errors on cardiac, the development reasoner 23	0 for every reasoner
Calibration conformance: a reported value must come from a measurement call in the reported convention	qwen reports the uncalibrated recomputed value, 6 silent (Table A8)	0; the rule fires on 26 sessions
Cross-path segmentation disagreement above 25% blocks the affected field and admits no argument	boundary kidney-erosion family, 5 to 8 of 8 silent	at most 1 of 8 for every reasoner
The call boundary fails closed on unknown argument keys	deepseek wraps arguments in an outer key in 62 sessions of the cardiac configuration	the call is refused and the budget is untouched
Value-level verification: a channel whose object is a value must have checked the value that is reported	The development reasoner checks the converted value but reports the original, on 1 of the 68 injected faults of the first additional CT sample	0 of 68; 17 repairs and 46.6% usable output against 20 and 48.3% without the rule
Input contract on the disagreement-summary tool: every value in one call must trace to the same field, and a mixed call is refused without executing	GPT-4o-mini passes values of five different organs into one call, so its alarm fires almost always and drives 60 of 100 natural-case downgrades	30 natural-case downgrades and 32% usable output on that reasoner, against 83 and 9% without the rule; no verdict of another reasoner on either volumetric task depends on this alarm
Primary-handle confirmation: a value from a substituted handle must be confirmed on the primary handle within ε_b	qwen and deepseek measure a blocked quantity on the second segmentation path and sign, 6 and 5 of 132 chest faults	0 of 132 for both; qwen keeps 62.3% usable output and 49 repairs with or without the rule, and deepseek has 56.3% and 46 against 56.7% and 47 without it

J SINGLE-FACTOR ABLATIONS OF THE ENFORCED RULES

Table A8 removes one enforced rule at a time from the full system and leaves the prompt, the tools and every other check untouched. Table A8b then extends the last row of Table A8, the only configuration that holds exactly the same information as the full system, to every task and reasoner.

Table A8b is the experiment that separates enforcement from information across the whole grid. With the evidence held fixed, the commit check accounts for 30 silent commits, 107 conflict releases and all six non-zero undischarged-alarm rates. These rates range from 0.7% to 19.1%, against zero on all twelve configurations of the full system (Table 2). With the protocol, usable output rises on both

Table A8: Single-factor ablations, one factor removed from the full system at a time. SE and CR count cases, out of 143, 131 and 132 injected faults and out of 80, 12 and 12 conflicts; Rep counts repaired cases over the same faults, U is the usable output rate and UAR the undischarged-alarm release rate. A rule that applies to a single task is reported on that task alone. Rows marked † were run on the injected-fault and conflict subset of cardiac MR, so their cardiac usable output and repairs are not comparable in scale with Table 1 and are omitted. The two abdominal CT rows with 63 repairs and 66.0% are separate runs under different protocol variants whose aggregate columns coincide; their accepted sessions differ on 24 cases. The reporting-handle row removes the per-value check on blocking items, the all-channels check and the primary-handle clause together, because the implementation controls them with a single switch. Table 4 of the main text groups the rows below by the component to which each belongs.

Configuration	Card. SE	CR	Rep	U %	UAR %	CT SE	CR	Rep	U %	UAR %	CXR SE	CR	Rep	U %
Full system	0	0	11	64.6	**0.0**	0	0	60	65.2	**0.0**	0	0	49	62.3
Witness grouping removed †	—	37	—	—	0.0	—	—	—	—	—	—	—	—	—
Calibration conformance removed	**6**	0	26	67.9	0.0	—	—	—	—	—	—	—	—	—
Provenance grounding removed †	7	0	—	—	0.0	—	—	—	—	—	—	—	—	—
Diagnostic-mode taint isolation removed †	0	0	—	—	0.0	—	—	—	—	—	—	—	—	—
Reporting-handle checks removed †	0	0	—	—	0.0	0	0	61	65.6	0.0	**6**	0	50	61.9
Fail-closed on undischarged items removed	—	—	—	—	—	0	0	63	66.0	**5.4**	0	0	48	62.7
Auxiliary channels admitted as witnesses	—	—	—	—	—	0	0	63	66.0	**5.4**	1	0	47	60.1
Self-comparison admitted as discharge	—	—	—	—	—	0	0	63	66.4	0.0	0	0	49	63.8
Discharge declarations removed	0	0	12	64.9	0.0	0	0	28	52.7	1.4	—	—	—	—
Contract only, no protocol (**information held fixed**)	0	40	25	56.8	**14.1**	0	0	62	64.8	**7.6**	4	3	37	52.6

Table A8b: The last row of Table A8 for every reasoner evaluated on each task, compared with the full system on the same cases. Both configurations hold the same channels, the same contract and the same alarms, and differ only in whether a program checks the commit. Counts are cases out of the injected faults, conflicts and distractors of each task.

Task	Reasoner	Contract only SE	CR	FA	U %	UAR %	Full system SE	CR	FA	U %	UAR %
Cardiac	qwen3.8-flash	0	40	39	56.8	14.1	0	0	0	64.6	**0.0**
Cardiac	deepseek-v4-flash	**14**	64	21	60.4	19.1	0	0	1	64.2	**0.0**
CT	qwen3.8-flash	0	0	0	64.8	7.6	0	0	0	65.2	**0.0**
CT	deepseek-v4-flash	0	0	2	66.4	15.8	0	0	0	66.0	**0.0**
Chest	qwen3.8-flash	**4**	3	4	52.6	2.7	0	0	4	62.3	**0.0**
Chest	deepseek-v4-flash	**12**	0	10	48.1	0.7	0	0	4	56.3	**0.0**

cardiac pairs, by 7.8 and 3.8 points, and on both chest pairs, by 9.7 and 8.2 points, and stays flat on abdominal CT.

The grid also shows why the tasks differ. On abdominal CT, the contract alone reaches zero silent errors with the strongest two reasoners, because the recomputation channel gives a clear signal on organ volumes and a competent reasoner acts on it. On cardiac MR, the same contract gives way twice, and it gives way most on the unarbitrable conflicts, with 40 and 64 releases of 80 against zero under the protocol. On chest radiography, it gives way on both reasoners, and three of the four wrong values that one of them signs are substitutions the protocol refuses. Abdominal CT alone would therefore understate what the commit check does, which is why we report the whole grid.

Four entries of Table A8 carry the rest of the argument. Removing the witness-group item raises cardiac conflict release from 0 of 80 to 37 of 80, while the prompt still states that copies of one record do not constitute a majority. Removing calibration conformance produces 6 silent errors, all along a single route: the reasoner recognises that the official value is suspect, adopts the uncalibrated value from the independent recomputation, checks that value against itself, and commits it as trusted with a note that the official number looks low. Removing the discharge declarations preserves safety but costs abdominal CT 12.5 points of usable output and 32 repairs; this is the measurement on which Section 3.3 rests. Removing the reporting-handle checks returns 6 of 132 chest faults to silence ($p = 0.031$, uncorrected) and adds no silent error on abdominal CT.

The UAR column tests Π_2 and the two rows that follow it in Table A8. Removing fail-closed on undischarged items raises it from zero to 5.4% on abdominal CT, and admitting advisory channels as discharge evidence has the same effect. Both routes around the obligation therefore become visible once the metric honours the declared scope. Silent errors stay at zero on these two rows, so the rules they remove do work that the SE column does not show: the values these configurations sign are correct, but their grounding is not established.

We removed fail-closed on undischarged items on abdominal CT and chest radiography but not on cardiac MR. The cardiac evidence for this factor is therefore Table A7a together with the first row of Table A7b. There, the same obligation stated as an instruction gave way 5 times on one reported reasoner and 23 times on the development reasoner, and never once when the program declared it.

A single channel used on its own as a rule marks the lower end of the same axis. Downgrading whenever the recomputation channel disagrees, with no other channel and no model in the loop, leaves 46% of cardiac faults, 48% of CT faults and 59% of chest faults silent with no false alarms. Its usable output is 66%, 49% and 51%, and its strict repair is 0%, 2% and 0%. Mask statistics used in the same way leave 64% and 60% of cardiac and CT faults silent, and on CT they also fire on half the distractors; physiological plausibility leaves 70% and 60% silent with no false alarms. Each channel covers the stage it audits and nothing else, which is the union of Section 3.2 reduced to a single term. At the other end of the axis is the rule controller, which runs every channel under the same downgrade-on-any-alarm policy. On cardiac MR it removes every silent error at 41% usable output, 24 points below our protocol. On chest radiography it falls 27 points below and still signs 4 wrong values. This is the trade-off that Table 1 shows our protocol avoids.

Table 2 also shows that grounding and correctness are separate properties. Without the contract and the protocol, the two weaker reasoners have the lowest undischarged-alarm rates of the four on abdominal CT and chest radiography, yet they have the most silent errors on every task: 63 of 143 on cardiac MR, 47 of 131 on abdominal CT and 96 of 132 on chest radiography. A model that disputes the alarms less often still signs wrong numbers. With the contract and the protocol in place, the same two models commit no silent errors on cardiac MR or abdominal CT.

Channel ablations remove the evidence and keep the rule. Removing the value-level hard evidence for the double-calibration family raises its silent count from 0.33 of 14, averaged over three runs, to 14 of 14. During these failures, the advisory multi-parameter channel did report differences above 15% on 11 of the 14 cases, and the reasoner attributed them to a calibration difference. Five families registered with a single covering channel keep zero silent errors when that channel is removed, owing to advisory redundancy, so the coverage matrix is conservative where it can afford to be. The design gives way where only defeasible evidence remains. This is the direction Proposition 1 states, and chest radiography measures it directly in its lateral-crop family.

K THE OPERATING ENVELOPE

Section 3.5 states as conditions the routes that Proposition 1 leaves open. This appendix reports where each route appears in the evaluation, either in the full system or under the ablation that reopens it.

A coverage hole requires a fault introduced at a stage that leaves no second trace. Two boundary families are registered as instances of this construction. The one-voxel erosion of the kidney mask turns out to leave a trace: without the cross-path rule of Table A7b, the reasoners commit 5 to 8 of its 8 cases. With a cross-path volume disagreement above 25% on one organ declared an undischargeable field-level item, every reasoner commits at most 1 of 8; in the remaining case the paths disagree by 16%, below the threshold. On the additional samples, the family appears in four cases whose two paths differ by 17% and 20%, also below the threshold, and these are the cases the reasoners miss there. On natural cases, the item fires in 4 of 100, and each of these four carries an error of 29% to 98% on the flagged organ against the AMOS22 expert annotation. The uniform 10-pixel dilation of the heart shadow on chest radiography leaves no such trace: the two segmentation paths shift together, the reported ratio stays plausible, and our system commits 6 of its 8 cases. Both families are registered outside coverage and scored in a separate column, and no result in this paper takes credit for either family.

An advisory-only family is closely related to the coverage hole, and the chest task measures it directly. The lateral field-of-view crop is covered by exactly one channel, the extent comparison. The design keeps this channel advisory because it fires on a correct value whenever the anatomy reaches the image border. The outcome therefore depends on whether the reasoner acts on a hint: the two primary reasoners silently sign none of the 12 cases, the two weaker reasoners sign 3 and 9, and the unprotected ReAct baseline signs 10. These 12 cases account for 12 of the 16 silent errors that our system commits across all tasks and reasoners. The other four come from Llama-3.3-70B. On three

heart leaks it measured only the primary path, so the cross-path item that carries the evidence never arose. On one photometric inversion, the metadata and appearance channels raised alarms without declaring a blocking item, which is a specification hole in the sense of Proposition 1. The criterion of Appendix I identifies this as an obligation that belongs in the program. Moving it there requires a blocking form of the extent comparison that separates the crop from the distractor family with the same signature; Section 5 names this as the next step.

A tolerance failure requires a fault smaller than the discharge tolerance of the item it triggers. With $\varepsilon_b = 1\%$ on the reported fields, a fault below this tolerance clears its item legitimately; this is the explicit trade-off between safety and usefulness. A tighter tolerance refuses more repairs. Table A8 reports the configuration in which the discharge declarations are removed altogether: on CT it preserves safety, while usable output falls from 65% to 53% and repairs fall from 60 cases to 28.

A witness failure requires copies of one record to pass as independent witnesses. The grouping of Π_3 closes this route: across three tasks and four reasoners, the 416 unarbitrable conflicts carry zero releases. What these zeros support is safety on those cases. A session with two witness groups does not contain the information needed to decide which group is stale. The honest outcome is therefore a downgraded verdict with the values reported, and resolving such a case requires evidence from outside the session.

Handle substitution is why the premise of Proposition 1 names the primary handle in addition to the reported one. In this route, the item is attached to the primary handle while the report names the second measurement path; the third clause of Π_2 then requires the value to be confirmed on the primary handle. Where the two paths agree within ε_b, the route is almost never attempted on the volumetric tasks. Where they disagree, removing the clause returns 6 of 132 chest faults to silence, with deviations of 4% to 15% from clean truth.

Grouping operates on the provenance that the task adapter models. Two declarations written by the same acquisition, sensor or vendor reconstruction share a cause. Work on multi-version programming showed how strongly such shared causes correlate failures: independently written implementations failed together far more often than independence would predict (Knight & Leveson, 1986). Groups therefore separate declarations by the origin that the adapter can trace. Extending the modelled provenance, for instance with a second measurement path from a different model family, extends what Π_3 can separate.

L Provenance of the baselines

Two rows of Table A9 are configurations we constructed ourselves. The rule controller implements the conservative policy that a safety-only metric rewards, and it is the reason Section 4.1 reports usable output beside every safety column. Appendix J evaluates it, and it appears in no main table. The contract-only configuration keeps the evidence and the contract of the full system but drops the commit check; it is the row of Table 3 that separates the effect of enforcement from that of information.

ToolGate, AgentSpec and RefineAct are not in this table. Each is a general framework whose specification comes from a policy, a trigger list or a user instruction, and none ships a medical measurement tool set. A quantitative comparison would therefore require porting each framework to this task and writing its specification ourselves, which would make the resulting baseline our own. That design space is represented by two rows that run author code: GuardAgent, a guardrail whose policy the model writes and the runtime executes, and Progent, a policy enforcer at the call layer.

M Replication and disclosure

Three additional samples outside the main benchmark test whether the results of Section 4 hold beyond its cases.

Under the same plan, we resampled and re-randomised the faults three times. The cardiac sample has 162 cases on 150 studies outside the main benchmark, with zero overlap at the study level. The first CT sample has 128 cases. Its 38 natural cases come from base studies outside the main benchmark, and its 90 injected cases are new pairings of a fault family with a base study. The base studies of

Table A9: Provenance of every baseline. Three baselines run author code and two are reimplemented from their papers. Every method uses the same reasoner as our system, and every baseline except ReAct receives the same channels, so the comparison isolates the commit decision. All agents except the ReAct baseline see alerts under the baseline contract view, which reports alerts without scope or discharge conditions, so the experiments measure the difference between an alert and a blocking item.

Method	Source	Implementation	Tool access	Tasks
ReAct	(Yao et al., 2023)	reimplemented from the paper	measurement tools of the task	all three
MedAgent-Pro	(Wang et al., 2026b)	reimplemented: planner, execute-and-assess, decider	same reasoner and channels as ours	all three
MedRAX	(Fallahpour et al., 2025)	author code, Apache-2.0, its loop	our channels injected as its tools	all three
Self-consistency	(Wang et al., 2023)	5 samples, majority over numeric equivalence	our channels	all three
LLM-as-judge	(Zheng et al., 2023)	second model reviews the session record	our channels	all three
GuardAgent	(Xiang et al., 2025)	author pipeline, our commit preconditions as its guardrail request	generates and runs check code	all three
Progent	(Shi et al., 2025)	author `secure_tool_wrapper`	model-written tool-call policies	all three
Rule controller	this paper	downgrade on any adjudicating alarm, no model	our channels	all three
Contract only, no protocol	this paper	ReAct prompt, full contract, no commit check	our channels	all three

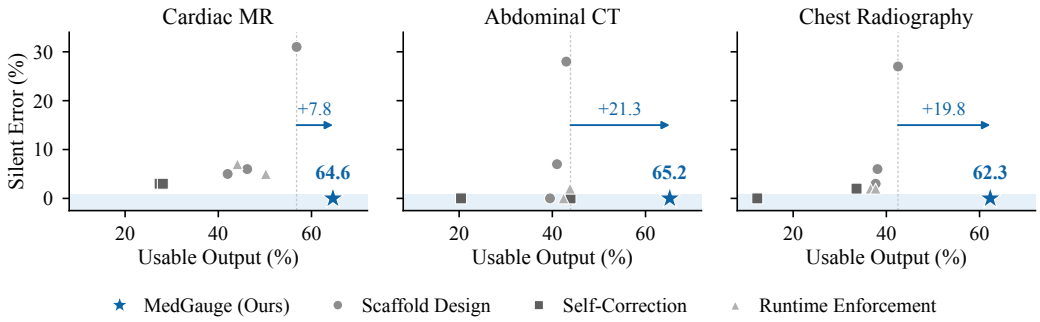


Figure A2: Usable output against silent error rate for the eight methods of Table 1, one panel per task, with `qwen3.8-flash` as the reasoner. The desired position for a measurement agent is the lower right. MedGauge reaches zero silent error and the highest usable output on every task. Each arrow starts at the usable output of the best baseline (dashed line) and marks the margin: 7.8, 21.3 and 19.8 points.

Table A10: Full per-task results of Table 1, adding the boundary family (Bnd: cases committed as trusted out of the 8 registered boundary cases), the undischarged-alarm release rate and the cost per case. All methods use `qwen3.8-flash`. ReAct sees only the measurement tools of the task and therefore has no UAR. On abdominal CT, the self-consistency, LLM-as-judge and GuardAgent runs did not generate the infeasible case, so their FA is computed over 23 distractors and their U over 255 cases.

Cardiac MR, 504 cases (144 injected faults, 143 of which move a reported value, 160 distractors, 80 conflicts, 120 natural)

Method	SE %	FA %	CR %	Rep %	U %	UAR %	¥/case
Scaffold design							
ReAct	31	18	82	2	56.8	n/a	0.027
MedAgent-Pro	6	39	85	0	46.2	24.9	0.038
MedRAX	5	47	74	1	42.0	24.1	0.039
Self-correction							
Self-consistency (k=5)	3	60	51	1	27.4	25.3	0.222
LLM-as-judge	3	68	33	2	28.1	18.0	0.064
Runtime enforcement							
GuardAgent	5	29	65	4	50.2	19.4	1.002
Progent	7	44	93	1	44.1	27.0	0.078
MedGauge (ours)	**0**	**0**	**0**	**8**	**64.6**	**0.0**	0.038

Abdominal CT, 276 cases (132 injected faults, 131 of which are feasible on their anatomy, 8 boundary, 24 distractors, 12 conflicts, 100 natural)

Method	SE %	Bnd	FA %	CR %	Rep %	U %	UAR %	¥/case
Scaffold design								
ReAct	28	0/8	4	100	3	43.0	n/a	0.027
MedAgent-Pro	7	1/8	4	0	0	41.0	3.5	0.044
MedRAX	0	0/8	12	0	0	39.5	13.7	0.051
Self-correction								
Self-consistency (k=5)	0	0/8	0	17	0	43.9	17.5	0.258
LLM-as-judge	0	0/8	70	0	1	20.4	7.7	0.068
Runtime enforcement								
GuardAgent	0	0/8	4	17	2	42.4	14.5	0.622
Progent	2	0/8	0	42	2	43.8	21.7	0.082
MedGauge (ours)	**0**	1/8	**0**	**0**	**46**	**65.2**	**0.0**	0.035

Chest radiography, 288 cases (132 injected faults, 8 boundary, 36 distractors, 12 conflicts, 100 natural)

Method	SE %	Bnd	FA %	CR %	Rep %	U %	UAR %	¥/case
Scaffold design								
ReAct	27	5/8	14	42	0	42.5	n/a	0.019
MedAgent-Pro	6	2/8	14	0	0	38.1	0.0	0.031
MedRAX	3	2/8	19	25	1	37.7	2.7	0.036
Self-correction								
Self-consistency (k=5)	0	0/8	61	8	0	12.3	2.9	0.206
LLM-as-judge	2	2/8	25	0	0	33.6	0.0	0.054
Runtime enforcement								
GuardAgent	2	1/8	22	33	1	36.6	3.8	0.342
Progent	2	2/8	22	50	0	37.7	5.3	0.086
MedGauge (ours)	**0**	6/8	11	**0**	**37**	**62.3**	**0.0**	0.032

these injected cases come from the pool of the main benchmark, because the AMOS pool available to us is exhausted at the study level. We added the value-level verification rule of Table A7b after the development reasoner failed on this sample, so the first CT sample is not held out for that rule; the cardiac sample and the second CT sample are. The second CT sample has 113 cases drawn under the same sampling rule. Its 90 injected cases are disjoint from all 294 (family, study) pairs of the main benchmark and the first sample, and its 23 natural cases come from the remaining unused base studies. The cardiac sample therefore tests generalisation across patients, and the CT samples test generalisation across fault instances.

Table A11: Additional samples outside the main benchmark. Counts are cases out of 66, 68 and 68 injected faults, 12, 6 and 6 conflicts, and 4 boundary cases per CT sample. Bnd counts boundary cases committed as trusted; these lie outside the covered families and are not counted in the SE column. All four reasoners of Table 2 run on the cardiac sample and the second CT sample, and a dash marks a sample on which a configuration was not run.

Configuration	Cardiac (162) SE	CR	U %	CT-1 (128) SE	Bnd	CR	U %	CT-2 (113) SE	Bnd	CR	U %
ReAct	28	10	47.3	19	0	6	33.1	—	—	—	—
ReAct + channels	3	11	46.7	0	0	2	28.8	—	—	—	—
MedGauge, qwen3.8-flash	0	0	54.0	0	1	0	56.8	0	2	0	55.3
MedGauge, deepseek-v4-flash	0	0	52.0	0	2	0	60.2	0	2	0	55.3
MedGauge, Llama-3.3-70B	0	0	2.0	—	—	—	—	0	1	0	24.3
MedGauge, GPT-4o-mini	0	0	5.3	—	—	—	—	0	2	0	15.5

The pattern of Table 1 reappears. The ReAct baseline is silent on 42% and 28% of faults, while our system commits none of the 672 covered fault-and-reasoner combinations and releases none of the conflicts ($p < 0.001$ against that baseline on both tasks). All four reasoners run on the cardiac sample and the second CT sample, and the two weaker models commit no silent error and release no conflict on these cases outside the main benchmark.

Natural cases, in which no fault was injected, show what the layer costs when nothing is wrong. With `qwen3.8-flash`, our system downgrades 14% to 16% of clean cases across the three tasks and reports their values. On chest radiography, that 14% is the lowest rate among the fifteen agent configurations, compared with 16% for the rule controller, 27% and 28% for the two configurations that hold the same channels without the protocol, and 81% for the sampling aggregate. On abdominal CT, the natural cases carry the AMOS22 expert annotations. The cross-path item fires on four of them, and in each the abstained organ has an error of 29% to 98% relative to the expert volume, so these abstentions mark real errors of the pipeline.

Natural cases against expert annotations. The natural cases of abdominal CT and chest radiography carry expert annotations: the AMOS22 organ labels and the SCR heart and lung contours. We score every primary field of all 200 cases against them, recomputing the reference from the original labels. Table A12 counts a field as a pipeline error when the audited pipeline deviates from the expert value by more than the task tolerance $\tau = 10\%$. It places the error in coverage when the second measurement path differs from the audited path by more than the declared threshold: 25% on CT, and on chest radiography 20% for heart width or 6% for thoracic width.

Table A12: Natural cases scored against expert annotations at $\tau = 10\%$. Signed counts the reasoner-and-field pairs, over the four reasoners of Table 2, that are committed as trusted with a reported value more than τ from the expert value; intercepted means abstained or downgraded. Fields counts the primary fields scored, and Errors counts the pipeline errors among them.

Task	Fields	Errors	In coverage	Signed in coverage	Outside coverage	Intercepted outside
Abdominal CT	500	62	8	0 of 32	54	5 to 15 of 54
Chest radiography	300	7	3	0 of 12	4	0 to 2 of 4

Within coverage, every error is intercepted under every reasoner, and in each CT session the log shows the path-disagreement item of `mask_stats` naming the field. None of the 44 reasoner-and-field pairs, which involve 11 distinct errors, is signed as trusted, and at $\tau = 25\%$ none of 32 is signed. Outside coverage the guarantee does not apply. On CT, 38 of the 54 errors outside coverage are shared by the two paths. Both paths come from one model family at two resolutions, so a systematic under-segmentation appears in both and their disagreement stays below the threshold. The direct way to turn such shared errors into path disagreement is a second measurement path from a different model family, such as an independently trained nnU-Net. Chest radiography has 4 errors outside coverage, too few to estimate a rate.

The same scoring tests the usable-output column, which counts a natural case as usable when it is trusted and every primary field is reported. Adding the requirement that every reported field lie within 10% of the expert value lowers usable output for every method, by 4 to 11 points on CT and by at most 1.5 points on chest radiography. Our system stays first on both tasks: its usable output falls

from 65.2% to 56.6% on CT and from 62.3% to 61.6% on chest radiography, and both values remain above the unadjusted value of every other method in Table 1.

N Extended related work

This appendix expands on the lines of work positioned in Section 2 and describes the specific mechanism of each.

Silent tool errors. When a calculator returns a wrong number without raising an error, models adopt the value and supply reasons for it (Sun et al., 2024); taxonomies organise such cases by the stage at which they arise (Ye et al., 2024). The closest existing responses remeasure the radiograph and rewrite the report (Heiman et al., 2025), flag implausible cardiac volume curves (Ruijsink et al., 2020), or estimate segmentation quality without ground truth (Cosarinsky et al., 2026; Marcianò et al., 2026). Each produces a new number or score, which is then accepted as the new truth. We run such estimators inside our observation layer as witnesses, and their alarms must themselves be discharged.

Scaffold design and self-correction. Both lines of work leave authority with the reasoner. Interleaving reasoning with tool calls is the common substrate (Yao et al., 2023), the design of the interface has a measurable effect (Yang et al., 2024; 2025), and medical agent frameworks fix the plan or the tool set (Wang et al., 2026b; Fallahpour et al., 2025). Self-correction checks the model's own output through sampling agreement (Wang et al., 2023), verbal reflection (Shinn et al., 2023), self-feedback (Madaan et al., 2023), tool-interactive critique (Gou et al., 2024) or independent re-answering (Dhuliawala et al., 2024), and it treats the tool as the reliable party. In these designs the instruction is the only remaining control, and GPT-4 with a safety prompt still fails 23.9% of emulated test cases (Ruan et al., 2024). These lines of work change what the agent can do; we change what it may certify.

Runtime enforcement. The line of work closest to ours gates actions with a program. Hoare contracts gate entry into a trusted symbolic state (Liu et al., 2026b), a policy language of triggers and enforced actions constrains the trajectory (Wang et al., 2026a), and RefineAct compiles the user's instruction to Prolog and queries it before each action (Batole et al., 2026). Other systems enforce least privilege (Shi et al., 2025), execute model-written policies (Xiang et al., 2025; Tsai & Bagdasarian, 2025), or supply rails, ontologies and isolation (Rebedea et al., 2023; Liu et al., 2026a; Wu et al., 2025). We share their view that the gate must be a program, and these gates already express rich conditions. AgentSpec evaluates predicates over the trajectory and can invoke a parameterised corrective action in place of the blocked one, and ToolGate lets a tool result update the trusted state only when its post-condition holds. The differences arise because the object we gate is a measurement. Their conditions are written as a policy, an instruction or a contract over the state that the agent exposes. Our conditions are computed by channels that exclude the stage they audit; they weigh a continuous quantity together with its measurement parameters and provenance, and they hold only within a registered coverage. After a violation, our agent is given the call that clears it, with the tool, parameters and tolerance declared in the blocking item. RefineAct, the closest of these systems, returns feedback on what was wrong with the blocked action. Section 4.4 approximates that feedback by keeping every alarm and removing its discharge condition; this is an internal proxy, and RefineAct itself is not run.

Measurement traceability. A learned predictor advises whether an echocardiographic frame can be measured reliably (Daghyani et al., 2026), and verifiable agents link final outputs to the intermediate artifacts and measurements behind them (Long et al., 2026). MedOpenClaw runs imaging agents inside a replayable viewer runtime. Once the episode ends, it scores the evidence they submit (key slices, coordinates and regions of interest) against withheld annotations (Shen et al., 2026). Traceability records where a number came from, so that a reader can act on it or an auditor can score it. Our blocking items, by contrast, hold a veto before the report is committed and are cleared by the evidence they name. The event we count follows the failure-detection literature, which names silent failure (Bungert et al., 2023) and formalises risk over undetected failures (Jaeger et al., 2023; Traub et al., 2024).

Channel independence has a long history in fault tolerance. N-version programming rests on the conjecture that independently written versions fail independently (Avizienis, 1985), yet versions written independently from one specification fail on the same inputs far more often than independence predicts (Knight & Leveson, 1986). Measurement admits a stronger form of independence, because

one physical quantity leaves traces of different origin in the data; Section 3.4 counts those traces by provenance.

Statistical and sampling-based filters for generated reports. ConfLVLM treats each claim of a vision-language model, including claims in radiology reports, as a hypothesis, and uses conformal prediction to bound the error rate of the retained claims under exchangeability (Li et al., 2025). Our protocol decides each case deterministically from the call record; its guarantee is conditional on coverage and states no population-level rate. RadFlag samples several reports at different temperatures and flags the claims that the samples do not consistently support (Zhang et al., 2025). It counts agreement among samples of one model, and these samples share that model's errors. For this reason, our witness groups count agreement only among sources of distinct provenance.