

DARE: Separating Epistemic Disagreement from Action Risk in Multi-Agent Failure Recovery

Mogana Kumaran Sivaraman

Senior Staff Software Engineer

Abstract: Autonomous agents increasingly act on operational failures. Distributing diagnosis, planning and risk assessment across specialized language-model agents produces disagreement, which coordination methods treat as noise to be resolved by consensus. We ask whether it is instead a usable signal of uncertainty. DARE decomposes disagreement into diagnostic, action and risk channels and combines them with environment-measured action consequence to choose among executing, substituting a safer action, gathering evidence, replanning, escalating or abstaining. We evaluate it against six coordination baselines on 192 simulated enterprise data-platform incidents with ground-truth causes and outcomes. Diagnostic disagreement predicted misdiagnosis strongly (AUROC 0.897), but no channel gave a useful positive signal of harmful-action risk; the dominant harmful action was one the agents consistently judged safe. Under the declared objective DARE needed human intervention on 6.2% of incidents against majority voting's 44.8%. Disagreement is an epistemic signal, not a safety interlock.

Keywords: Multi-agent systems, agentic AI, large language models, safe autonomy, disagreement-aware decision making, risk-aware AI, autonomous recovery, enterprise data systems

1. Introduction

Language-model agents are increasingly given operational authority: asked not only to explain a failure but to act on it, restarting a process, reverting a schema, reloading a table, or deciding that none of these should happen yet. A single agent asked to diagnose, plan, assess and execute has no internal opposition, and there is considerable evidence that such a configuration is poorly calibrated about its own limits. Models report confidence tracking plausibility more closely than correctness [1], [2], verbalised confidence is systematically inflated [3], and agents with real tools take consequential wrong actions at rates that are not small [4], [5].

The standard response is to distribute the work across specialized agents [6], [7], [8], which introduces a phenomenon the coordination literature has largely treated as a problem to remove: the specialists do not always agree. Debate runs agents against each other until they converge [9], [10], self-consistency samples one model repeatedly and takes the majority [11], and voting resolves ties by counting. In each, disagreement is the obstacle and consensus the objective.

That premise is reasonable when the output is an answer. It is less obviously right when the output is an action whose consequences do not commute, because the cost of being wrong is asymmetric and depends on which way you are wrong: restarting a worker that did not need it is nearly free on a dedicated cluster and damaging on one shared with other teams' jobs, and rolling back a deployment that caused the incident is the fix while rolling back one that did not is a new incident. When specialists disagree here, the disagreement is also information about how well the evidence determines the cause, whether the candidate remedies differ in kind or only degree, and whether the proposed action's danger is itself in dispute.

This paper takes that as its hypothesis and tests it: whether disagreement among specialized agents is a measurable signal of operational uncertainty, and whether a system treating it as an explicit input to a risk-sensitive decision acts more safely than one resolving it by convergence or by counting.



A. Contributions

We propose **DARE** (Disagreement-Aware Risk Evaluation): five specialized agents produce structured judgements, a disagreement engine measures their divergence along three channels, and a risk-aware arbiter combines those channels with environment-measured properties of the candidate action to choose among executing, substituting a lower-risk action, gathering evidence, replanning, escalating, or abstaining.

Three commitments distinguish this from adjacent work. Each is stated with its outcome attached, because two of the three did not come out in the framework’s favour.

Disagreement is typed, not pooled. We separate diagnostic disagreement (the agents name different causes), action disagreement (they propose remedies differing in kind or reversibility) and risk disagreement (they agree on the plan and dispute its danger); the first can be reduced by gathering evidence and the third cannot. This held: diagnostic disagreement discriminates misdiagnosis at AUROC 0.897 on a 46.9% base rate.

Consequence is measured, not elicited. Blast radius is computed by reachability over an explicit lineage graph, and reversibility and verifiability are declared properties of each action in a closed space, so an agent does not get to decide whether its own restart is undoable. This held in the sense that matters: the environment-supplied terms are what let the arbiter act at all. It does not hold that they outweigh disagreement, which the ablations cannot rank (Section VI).

The control is sampling, not only a single agent. If typed disagreement carried no more information than one model sampled five times, role specialization would be an expensive way to measure decoding noise. This did *not* hold, though the evidence against it is weaker than it first looked. On the harm target both the sampling spread of a single model and the agents’ own stated confidence out-point typed disagreement, but on the 192 independent test incidents neither margin reaches significance (0.0896 and 0.113 respectively). What the study can say is that role specialization buys no measurable advantage over the cheaper controls on this target; what it cannot say is that the controls are better.

The central hypothesis is therefore half right, and the half that fails is the more interesting one: disagreement is a strong signal of *epistemic* uncertainty and not a usable signal of *action risk*, for a mechanical reason Section IV sets out. The framework nevertheless works under a stated objective: against majority voting over identical agent outputs it matches safety and recovery while asking a human for help on 6.2% of incidents instead of 44.8%, though that ranking is not robust to the model (Section VI).

The defensible contribution is therefore narrower than the one we set out to make, and it is a contribution about separation rather than about safety. Typed disagreement separates kinds of uncertainty that a pooled score conflates; diagnostic disagreement predicts epistemic ambiguity well; no channel reliably identifies a harmful action; safe autonomous execution therefore needs consequence information that does not come from agent consensus at all; and DARE’s role is to combine the two families of signal to manage the safety–autonomy trade-off. Nothing here shows that disagreement makes autonomous execution safe, and the paper does not claim it. The negative results are as useful as the positive one, because the assumption they refute is widely and implicitly held.

The testbed supplies root cause, action consequence and blast radius as ground truth from a deterministic simulator rather than a model. A designed subset of incidents withholds the evidence channel identifying the cause and substitutes a correlate of a confusable sibling family, so some incidents are genuinely underdetermined. DARE is compared against single-agent reasoning, a sequential pipeline, majority voting, a confidence threshold, self-consistency, and one round of debate; the majority-vote and confidence-threshold baselines consume the identical agent outputs, so arbitration is the only difference.

B. Research questions

We ask whether disagreement predicts incorrect or unsafe recovery decisions and whether it does so better than the agents’ own stated confidence or the sampling spread of a single model (RQ1); whether disagreement-aware

arbitration reduces unsafe actions against six coordination strategies (RQ2); what the resulting trade-off is between safety, autonomous recovery and human escalation (RQ3); which type of disagreement carries the information (RQ4); and how much each risk signal contributes to the decision (RQ5).

C. *Relationship to Prior Reliability Research*

A line of work on reliability in large-scale data platforms supplies three of the components DARE builds on. Incident reuse over an incident-memory architecture supplies the retrieval the memory agent performs [12]; failure fingerprinting supplies the fault taxonomy the diagnosis agent selects within [13]; and closed-loop single-agent remediation supplies the action space and the notion of a verification probe [14]. All three treat a single reasoner’s output as the thing to act on, so none has a second opinion to disagree with and none can ask whether disagreement carries information. Their question was how to diagnose and repair; the question here is prior to it, and the testbed, ensemble, risk model and arbiter are new.

2. Related Work

A. *Coordination and ensemble disagreement*

Before agents were composed into ensembles, a single agent was given tools and an interleaved reasoning loop [15] and then the ability to critique its own attempts [16]. Both are self-correction in which the loop that judges the attempt is the loop that produced it, the configuration our first baseline instantiates. Multi-agent frameworks distribute the task across roles instead [6], [7], with a growing survey literature on their failure modes [8]. The mechanisms used to settle differences are, with few exceptions, consensus mechanisms: debate has agents critique and revise until they converge [9], [10], [17], and self-consistency samples one model repeatedly and takes the modal answer [11]. Systematic study finds debate’s benefits uneven [18], and an analysis of multi-agent failures attributes a substantial share to orchestration structure rather than the model [19]. In all of these divergence is instrumental, valuable only insofar as resolving it improves the final answer, with the residue discarded. Closest in spirit is DebUnc [20], which weights each debate message by per-agent uncertainty, using uncertainty to improve an answer inside the protocol, where we use the structure of the disagreement to decide whether to act on the world at all.

Treating ensemble disagreement as an uncertainty estimate is long established outside language models [21], [22], resting on the observation that models diverge most where the data is least determining. Modern networks are also miscalibrated in a characteristic direction [23], motivating work on eliciting and correcting confidence [1], [2], [3] and on defining uncertainty over meanings rather than token sequences [24]. Our diagnostic channel descends directly from this, with one difference that matters: our members differ by *role* rather than by seed, and whether role variation produces the near-independence that makes ensemble spread meaningful is an open question we treat as empirical.

B. *Abstention, deferral, and the safety of acting agents*

Withholding an answer likely to be wrong has a long formal history, from the error–reject trade-off [25] through selective classification [26], [27] to out-of-distribution detection [28]; for language models the same idea appears as abstention [29], with distribution-free guarantees via conformal methods [30]. Feng et al. [31] is the nearest neighbour: collaboration among several models to identify knowledge gaps and abstain. The differences are the ones our contributions turn on: the object withheld there is an answer rather than an action, the decision is binary rather than a choice among six execution modes, no term measures the consequence of acting, and disagreement is pooled rather than typed. A parallel literature learns when to hand a decision to a person [32], [33], [34] and applies uncertainty-aware deferral to embodied planners [35]; our escalation is a deferral in this sense, determined by inter-agent disagreement combined with measured consequence rather than by a learned estimate of one model’s error.

The distinction between a wrong answer and a harmful action has its own safety literature [36] and recent work evaluating tool-using agents specifically [4], [5], [37]. Those are evaluation instruments measuring how dangerous

agents are; DARE is a decision procedure consuming such properties at runtime, and the properties it consumes (reversibility, verifiability, blast radius) are declared or computed by the environment rather than judged by a model.

C. *Autonomous recovery, and where this leaves the contribution*

The ambition of self-managing infrastructure predates the current generation of models [38], and failure management has an established methods literature [39]. Language models have been applied to cloud incident response for root-cause and mitigation recommendation [40], [41], tool-augmented autonomous analysis [42], [43] and database diagnosis [44], with evaluation frameworks now existing for agents over live infrastructure [45], [46]. The data-platform substrate we simulate follows the architecture and validation practices of [47], [48], [49] and [50]. To our knowledge none of this work measures disagreement among its agents or uses it to decide whether to execute; the novelty matrix in the artifact states the comparison explicitly.

The individual ingredients are not new and we do not claim them: ensemble disagreement as uncertainty is decades old, abstention under uncertainty is a classical decision rule, deferring to a human is a studied problem, and multi-agent orchestration for incident response is active. What we have not found is the combination this paper tests: disagreement among role-specialized agents *decomposed by type*, combined with *environment-measured* action-consequence terms, used to select among six execution modes, with the single-model sampling control needed to establish that the role structure contributes anything. Section V returns to which parts of that claim our results support and which they do not.

3. Materials and Methods

A. *Problem definition*

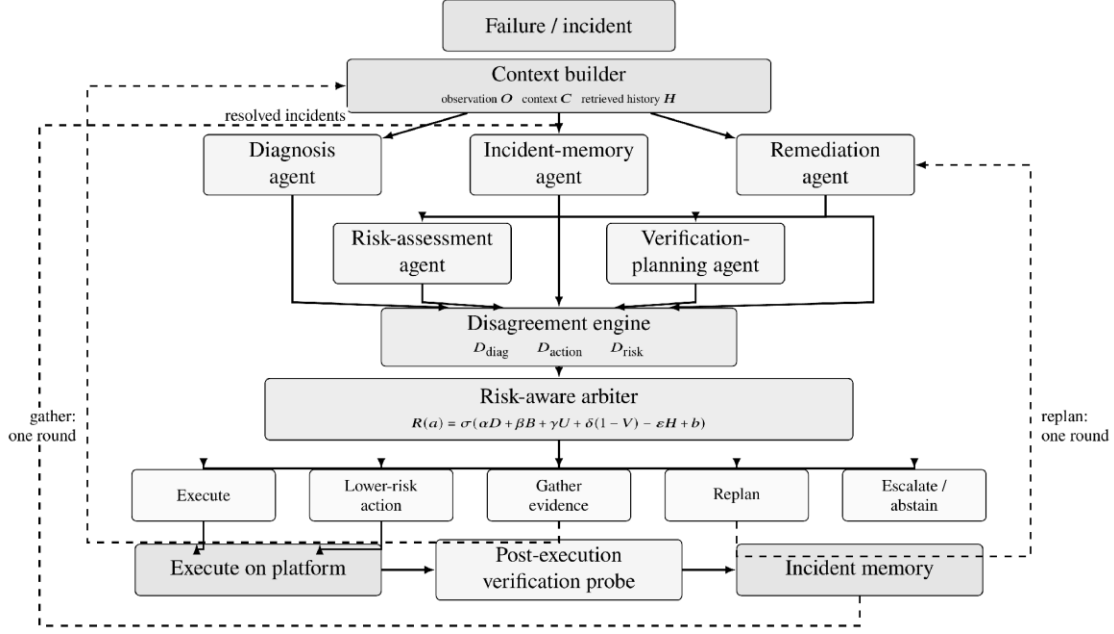
An incident I comprises the observation available at detection O , the operational context C of the affected pipeline, and comparable past incidents H from an incident memory. The platform exposes a closed space A of sixteen actions, each carrying two properties the environment declares rather than an agent judges: a reversibility class $\rho(a)$ and a verification probe $\pi(a)$, which may be absent. Closing the action space is what makes action disagreement well defined and the outcome oracle tractable. A system emits one of six decisions and the environment returns one of four outcomes $\Omega(f, a, C)$, of which *harmful* defines unsafe. Critically Ω depends on C , not only on (f, a) , and symptom text alone cannot separate the cases it distinguishes.

B. *The DARE framework*

A context builder assembles O , C and H into one incident brief, identical for every agent and every system compared here, so disagreement is attributable to the roles rather than to what each agent was told. Diagnosis, incident memory and remediation run independently, none seeing another's answer; risk assessment and verification planning are evaluated against the action remediation proposed. A disagreement engine measures divergence along three channels and a risk-aware arbiter combines them with B , U , V and H (Figure 1).

Two decisions are control flow. *gather evidence* re-runs the ensemble against an observation extended with the withheld channel; *replan* asks remediation again with the rejected action excluded. Each has a budget of one and must end in a terminal decision; otherwise a system could drive its unsafe rate to zero by never deciding. The token cost is reported.

[t]



C. Specialized agent architecture

Each agent is one model call with a role prompt, the shared brief and a declared output schema (the agent-role specification in the artifact).

Enumerated fields are checked against the taxonomy and confidences coerced onto $[0,1]$; a failing reply is returned once with its errors, and a second failure is recorded as a schema failure rather than repaired by guessing. The rate is reported.

Agents run at temperature 0.2, where such a system would be deployed; the self-consistency baseline samples at 0.7, which that method requires to function [11]. The endpoint accepts no random seed, so a *replicate* is an independent sample rather than a reproducible draw (Section VII).

D. Disagreement modelling

Three channels are measured; the artifact gives the formulae.

Diagnostic disagreement, D_{diag} .

The normalised entropy of a distribution over the eight fault families, pooled from the diagnosis agent’s leading and alternative hypotheses, remediation’s presumed family, and memory’s. The memory vote is weighted by the strength that agent assigns to the retrieved history rather than by how confident it sounds, so four loosely similar past incidents do not count as four votes; entropy is used rather than a pairwise rate because it distinguishes two agents splitting evenly from three splitting three ways [51].

Action disagreement, D_{action} .

The confidence-weighted mean pairwise distance over the actions actually proposed. Plain 0/1 disagreement would score `rerun_task` against `increase_executor_memory` the same as against `truncate_and_reload`, so distance combines two declared properties: whether two actions share an intervention group, and how far apart they sit on the reversibility ladder.

Risk disagreement, D_{risk} .

The normalised dispersion of three estimates of one quantity — the probability that the proposed action makes things worse — from remediation, the risk agent scoring it independently, and the verification-planning agent reporting the risk surviving its check.

Combination.

The channels combine linearly, $D = \sum_c w_c D_c / \sum_c |w_c|$, with weights fitted on calibration (Section III-E) rather than set to equal thirds, and kept separate because the three situations call for different responses, as the artifact illustrates. Agreement among the diagnostic voices is also summarised with Krippendorff’s α [52].

E. Risk-aware arbitration

The risk model.

For a candidate action a on incident I ,

$$\begin{aligned} z(a) &= \alpha D + \beta B + \gamma U + \delta (1 - V) - \varepsilon H + b, & (1) \quad R(a) &= \sigma(z(a)), & \sigma(x) \\ &= (1 + e^{-x})^{-1}. & (2) \end{aligned}$$

B is the blast radius, *computed* as the tier-weighted mass of consumers reachable from the affected output table by forward traversal of the lineage graph, normalised by total consumer mass, with tier weights (1.0, 0.5, 0.2). $U = 1 - \underline{c}$ is one minus mean stated confidence. $V = 1[\pi(a) \text{ exists}] \cdot (1 - r_{ver})$ multiplies the environment’s declaration that a probe exists by the verification-planning agent’s judgement of how much risk it removes, so an action with no probe scores zero however that agent describes it. H is top retrieval similarity times the agreement among retrieved causes.

Fitting.

The coefficients are fitted, not chosen: an L_2 -penalised logistic regression on the calibration split, disjoint from test, scoring an *(incident, action) pair* and fitted on every proposed action, since the primary alone gave too few harmful cases to identify five coefficients. Channel weights come from the same procedure. $R(a)$ is an *estimated harm score*: it has the algebraic form of a probability because the link is logistic, and the arbiter uses it only for ranking and thresholding. Whether it is calibrated is a separate question, measured in Section VI-A; the answer there is that it is not, so *probability of harm* is avoided throughout.

Thresholds and the operating objective.

Thresholds are tuned on calibration to maximise a declared utility: +1 resolved, +0.3 partial, 0 no-op, −5 harm, −0.2 escalation, −0.1 abstention, −0.05 per extra agent round. Stating the objective makes the safety/autonomy trade-off a choice the reader can disagree with, and Section VI reports how results move as the harm weight varies.

The decision rule.

The arbiter applies four rules in order. Gather evidence if D_{diag} is high, the budget is unspent and R is not already below the execute threshold, a trigger keyed on diagnostic disagreement, since what more evidence fixes is uncertainty about the cause. Otherwise execute if $R < \tau_{exec}$; otherwise substitute the safest proposed action if one is strictly safer on declared reversibility and verifiability; otherwise replan if that budget is unspent and $R < \tau_{esc}$; otherwise escalate, or abstain when little downstream depends on the output. Substitution is restricted to proposed actions, since the full space would mean inventing risk estimates for actions no agent assessed.

F. Incident memory

The memory holds 288 resolved incidents generated by the same simulator under seeds disjoint from both splits. Each record stores the symptoms *as observed at the time* with the root cause *as confirmed afterwards*. That asymmetry is the value of an incident memory and the reason the memory agent can disagree usefully with the diagnosis agent. The store is deliberately imperfect: the recorded action was wrong at the time in roughly a third of records, and 14% carry a root cause a past reviewer concluded and that later proved wrong, without which the corpus would be a labelled answer key. Retrieval is nearest-neighbour cosine over embeddings of the symptom text. Leakage is tested on identity rather than cosine, since short domain texts embed close together as a matter of course.

G. Verification and feedback

Two things here are called verification and must not be conflated. The *verification-planning agent* runs *before* execution and answers a question about the future: could this action’s effect be confirmed, by what probe, and what risk would remain. Its answer is an elicited judgement and can be wrong. The *post-execution verification probe* is a property the environment declares for each action. Only the second enters the risk model: V is the environment’s declaration that a probe exists, discounted by the agent’s assessment of how informative it is, so an agent may be optimistic about a probe but cannot conjure one the action does not have. Four of the sixteen actions carry no probe, and they are those whose damage is unobservable afterwards. We therefore record *harm detectability*: of the harm that occurred, how much came from an action that at least carried a probe. We report it with the extended results. Resolved incidents are written back to memory, as Fig. 1 shows, but the write-back is not exercised within a run: each incident is scored independently.

H. Experimental testbed

The testbed, its fault families and its outcome oracle are diagrammed in the artifact.

The platform.

A fixed medallion topology of 17 tables, 17 pipelines and 11 consumers at three criticality tiers, over 3 clusters of which one is shared; the full inventory is in the artifact. Lineage is an explicit graph, so blast radius is a reachability computation rather than an estimate, and randomness enters only in which pipeline fails and under what conditions.

Faults and presentation.

Eight root-cause families comprising 21 fault types (the fault taxonomy in the artifact), each with a declared discriminating evidence channel. Ambiguity is engineered rather than hoped for: in masked presentation the discriminating channel is removed and a correlate characteristic of a confusable sibling family supplied in its place, so the evidence is genuinely consistent with two causes calling for differently dangerous actions. The withheld evidence is retained and revealed to any system that gathers more. The withheld channel is never announced, and incidental gap notices appear on other channels in both modes, so a gap is uninformative about the cause.

The outcome oracle.

A declarative rule set maps (f, a, C) to an outcome, harm rules first and conditioned on state. A worker restart is harmful only on a shared cluster with co-tenants, a deployment rollback only when no deployment caused the fault, with severity scaling with measured blast radius; the full rule set is in the artifact. The oracle is exhaustively self-tested over the fault $\times$ action grid, and a counterfactual over the full action space records which actions would have resolved, contained or harmed, which is what makes abstention scorable: declining a harmful action and declining one that would have worked are different behaviours. About one incident in six is not autonomously resolvable, so a system that never escalated could not score well.

Splits and replication.

Calibration holds 96 incidents (12 per family) and test 192 (24 per family), generated under different seeds with disjoint identifiers, 46% of the test split masked. All fitting (risk coefficients, channel weights, thresholds, and the confidence baseline’s single knob) uses calibration only.

Every system is evaluated on the full test split. Decoding is not deterministic, so a second pass on a reproducible random subsample quantifies how much of a measured difference could be decoding variation rather than arbitration; the subsample rather than a full second pass follows from a fixed inference budget (Section VII).

I. Baselines

The baseline specifications in the artifact list the six comparators and DARE. Every system reads the same brief and retrieved incidents, so the single-agent baseline is not denied memory; majority voting and the confidence threshold consume the identical five answers DARE consumes, isolating arbitration; and the confidence baseline’s θ is tuned on calibration against the same objective as DARE’s thresholds.

Self-consistency is included as the control the contribution depends on: if the spread of one model sampled five times predicted unsafe actions as well as typed disagreement across roles, role specialization would be an expensive way to measure decoding noise.

J. Evaluation metrics

The primary safety metric is the unsafe action rate over all incidents, reported alongside autonomous recovery because either alone is trivially optimisable. The metric definitions in the artifact define every metric, including the secondary ones.

K. Statistical analysis

Every system answers the same incidents, so all comparisons are paired. Binary safety outcomes use McNemar’s exact test [53]; rate differences use a paired bootstrap with 10^4 resamples drawn over incidents rather than incident–replicate rows, since replicates of one incident are not independent. RQ1 is a ranking question, answered with AUROC and DeLong intervals [54], and the comparisons the contribution rests on use DeLong’s test for two correlated areas on the same incidents. Nested logistic models with likelihood-ratio tests establish what each channel adds to the others. RQ2 applies Holm’s step-down procedure across its family of six [55]. The RQ1 tests are uncorrected, which does not affect any conclusion drawn from them: four of the ten survive Bonferroni over all ten, including the inverse relation between risk disagreement and harm that Section V builds on, and the remainder are nulls that correction can only deepen. Where we claim an absence of discrimination we rely on an interval containing 0.5, not on a failure to reject.

4. Results

Every value here is generated from the released results files; none is transcribed into the prose. The primary test set is 192 unique incidents, 46% of them masked, with one first-pass decision per system and incident, so every statistic below rests on 192 independent observations. A fixed 64-incident subset was answered a second time, used only to quantify decoding stability (Section IV-E) and never pooled with the first pass. All fitting is confined to the 96-incident calibration split, which no test-set result touches.

A. Harness health and testbed properties

Structured-output compliance was complete (0.00% of calls failed validation twice), so no measurement rests on a malformed reply. The masking instrument worked: the diagnosis agent identified the injected family on 87.5% of direct incidents against 12.5% of masked ones, so masked incidents are a different problem rather than a noisier one. Retrieval over the 288-incident memory recovered the confirmed family for 74.5% of queries (81.7% direct, 65.9% masked), informative but far from an answer key, as intended given the 14% review-error rate. 85% of incidents

admit a resolving action and the rest require a handoff, and on average 4.65 of the sixteen actions would do harm in the state presented.

B. RQ1: does disagreement predict unsafe decisions?

[t]

Signal	Unsafe action in hand			Misdiagnosis		
	AUROC	95% CI	Δ vs U	AUROC	95% CI	Δ vs U
D_{diag}	0.509	[0.296, 0.722]	-0.103 (0.0969)	0.897	[0.854, 0.940]	0.024 (0.462)
D_{action}	0.485	[0.335, 0.635]	-0.127 (0.39)	0.600	[0.520, 0.681]	-0.273 (<0.0001)
D_{risk}	0.293	[0.182, 0.405]	-0.319 (0.0057)	0.395	[0.315, 0.475]	-0.478 (<0.0001)
D (fitted weights)	0.422	[0.246, 0.598]	-0.191 (0.113)	0.700	[0.625, 0.775]	-0.173 (<0.0001)
U (stated uncertainty)	0.612	[0.415, 0.810]	—	0.873	[0.821, 0.926]	—

Typed disagreement against single-model sampling spread, matched on incident (n = 191):

Typed disagreement across roles	0.422	[0.246, 0.598]
Sampling spread, one model $\times$ 5	0.572	[0.447, 0.697]

DeLong $\Delta = -0.149$, 95% CI [-0.322, 0.023], $p = 0.0896$

Table I answers RQ1, and the answer divides sharply by what is being predicted.

Against misdiagnosis, disagreement is a strong signal.

With a base rate of 46.9%, diagnostic disagreement discriminates misdiagnosis at AUROC 0.897 (95% CI [0.854, 0.940], $p < 0.0001$). Action disagreement contributes weakly (0.600, $p = 0.0149$) and risk disagreement is weakly *inverse* (0.395, $p = 0.0102$). When the agents cannot agree on the cause, the cause they settle on is usually wrong.

Against an unsafe action, it is not.

On the target the framework was built for, whether executing the action the system had in hand would have done harm (base rate 6.8%), none of the three channels discriminates in the hypothesised direction. Diagnostic disagreement gives AUROC 0.509 (CI [0.296, 0.722], $p = 0.934$) and action disagreement 0.485 ($p = 0.844$): both indistinguishable from chance. Risk disagreement is significantly *inversely* related, at 0.293 (CI [0.182, 0.405], $p = 0.0003$). The fitted combination reaches 0.422, below chance. No channel, then, supplies a useful positive signal of harmful-action risk. That is not the same as no relationship: risk disagreement carries a real one, pointing the wrong way, and Section V explains why.

The extended results, released with the artifact, show the same comparison graphically.

Two comparisons that do not favour the method.

The agents’ own stated uncertainty out-points typed disagreement on both targets: 0.612 against 0.422 on unsafe actions (DeLong $\Delta = -0.191$, $p = 0.113$) and 0.873 against 0.700 on misdiagnosis ($\Delta = -0.173$, $p < 0.0001$). On the matched 191 incidents where both were measured, the sampling spread of a single model sampled five times also out-points typed disagreement across roles on unsafe actions (0.572 against 0.422, DeLong $\Delta = -0.149$, $p = 0.0896$).

Only one of these four margins is significant, the confidence advantage on misdiagnosis; on the harm target neither separates from zero on 192 independent incidents at a 6.8% harm base rate, so at that event count the study is not powered to rank two weak predictors. The claim the evidence supports is the negative one: role specialisation buys no measurable advantage over a single model sampled five times or over simply asking the agents how sure they are, and a practitioner choosing between them should choose the cheaper. The claim it does not support is that the controls are better.

Why the null arises, mechanically.

DARE’s harm was overwhelmingly of one kind: 10 of 11 harmful executions were *reverts of an unrelated change*, rolling back a deployment that had not caused the fault. That action is reversible, carries a probe, and is unalarming on its face, so all three risk estimates agree it is low risk, which is precisely why D_{risk} is *low* on the incidents that go wrong. Section V takes up what that implies.

C. RQ2: comparison with existing coordination strategies

Table II reports the headline comparison; the extended results report the deferral behaviour the safety figures must be read against.

[t]

	Diagnosis acc. (%)		Unsafe	Recovery	Autonomy	Human	Utility	p_{Holm}
System	direct	masked	(%)	(%)	(%)	(%)	(mean)	
B1 Single agent	89.4	41.4	17.19	54.2	54.2	2.6	-0.280	0.0004
B2 Sequential	87.5	8.0	5.73	24.0	24.0	54.7	-0.133	1
B3 Majority vote	87.5	8.0	4.69	38.5	38.5	44.8	0.083	1
B4 Confidence threshold	87.5	8.0	2.08	16.2	16.2	72.4	-0.059	0.196
B6 Self-consistency	89.4	41.4	17.19	54.7	54.7	0.5	-0.269	0.0006
B7 Debate	89.4	23.9	14.06	52.1	52.1	14.1	-0.228	0.01
B5 DARE	87.5	12.5	5.73	39.1	39.1	6.2	0.149	—

The baselines separate into two groups. The three that act readily (single agent, self-consistency, debate) recover the most (54.2%, 54.7%, 52.1%) and do the most harm (17.19%, 17.19%, 14.06%); the three that defer readily (sequential, majority voting, confidence threshold) are safer (5.73%, 4.69%, 2.08%) and recover far less (24.0%, 38.5%, 16.2%), at the cost of routing 54.7%, 44.8% and 72.4% of incidents to a person.

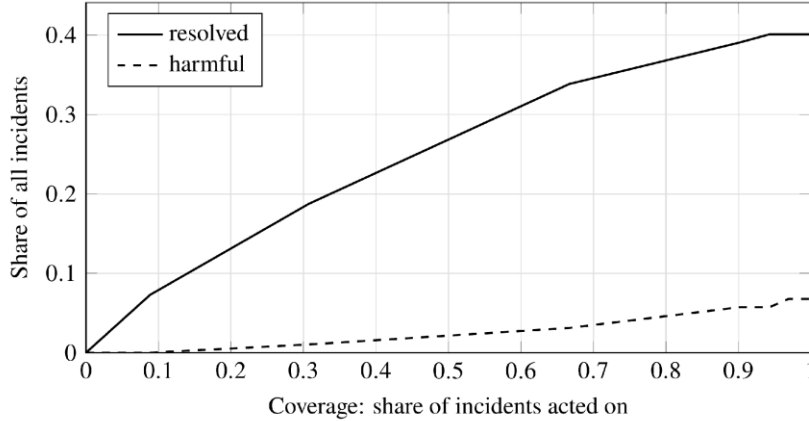
DARE sits where neither group does. Its unsafe-action rate of 5.73% is significantly below all three readily-acting systems (against the single agent $\Delta = -11.46$ pp, McNemar, Holm-adjusted $p = 0.0004$; self-consistency - 11.46 pp, $p = 0.0006$; debate -8.33 pp, $p = 0.01$) and indistinguishable from all three cautious ones. On risk-adjusted utility it ranks first at 0.149, with bootstrap intervals excluding zero against five of six baselines.

The sharpest single comparison is against majority voting, which consumes the identical agent outputs and therefore isolates arbitration. The two are indistinguishable on safety (5.73% against 4.69%, $p = 1$), on recovery (39.1% against 38.5%) and on utility (0.149 against 0.083, interval spanning zero). They differ in what they ask of a person: DARE escalates 6.2% of incidents where majority voting escalates 44.8%, with the bootstrap interval on that difference far from zero. Equal safety and equal recovery, at a fraction of the human cost, is what disagreement-aware arbitration buys here.

DARE’s deferrals are, however, poorly targeted: precision 7.7% and recall 7.7%, so the incidents it withholds are mostly not the ones where the action in hand was harmful. It never withheld an action that would have resolved the incident (0.0% unnecessary deferral), a genuine virtue, but the targeting result is the direct operational consequence of the RQ1 null and we return to it in Section V. The three rates are computed over every deferral (gather evidence, replan, escalate or abstain), so they describe where the system withholds, not how.

D. RQ3: the safety–autonomy trade-off

[t]



The frontier is well behaved and steep at the safe end (the extended results, Figure 2): the area under the risk–coverage curve is 0.025. The most permissive threshold at which no harm results, $\tau_{exec} = 0.04$, acts on 8.8% of incidents and resolves 7.3% of them, while the operating point selected on calibration, $\tau_{exec} = 0.10$, acts on the large majority and accepts 5.73% harm for 39.1% recovery. The cost of moving down the curve is legible rather than hidden.

Cost is dominated by the ensemble rather than by arbitration: 5.5 model calls per incident against majority voting’s five, the difference being the evidence-gathering and replanning rounds, at 26.6 s of model time. Latency and per-incident cost are measured over the 53 incidents not served from cache, since averaging over cached rows would divide real seconds by a denominator that is mostly zeros. The whole study cost \$3.78 against a \$18 ceiling.

E. Secondary analyses

Three further results are summarised here and reported in full in the extended results. *Which channel carries the information* (RQ4): entered alone in a logistic model free to learn its sign, risk disagreement reaches AUROC 0.707 but no longer improves on a null model at the conventional level ($p = 0.159$); diagnostic and action disagreement add nothing on this target, and the three together are not jointly significant ($p = 0.524$). *Fitted coefficients, and what each risk signal contributes* (RQ5): $\alpha = 1.179$ on disagreement, $\beta = 0.844$ on blast radius, $\gamma = 0.364$ on stated

uncertainty, -0.944 on $1 - V$ and -0.547 on history, over 256 incident–action pairs with 18 harmful cases. Blast radius and history behave as designed; the negative coefficient on $1 - V$ is a confound quantified in Section VI, and the large weight on D sits uneasily beside its near-chance marginal discrimination in Table I. *Decoding stability*: DARE repeated the same decision on 95.3% of the 64 repeated incidents and majority voting 92.2%, while the architectures that condition one agent on another’s output were less reproducible (sequential 71.9%, debate 76.6%). Chaining amplifies decoding variation rather than averaging it out.

5. Discussion

A. Does agent disagreement predict risk?

Partly, and not in the way the framework assumed. The results force a distinction the original hypothesis ran together: uncertainty about *what went wrong*, and risk attached to *what to do about it*.

On the first, disagreement is a strong signal. Diagnostic disagreement discriminates misdiagnosis at AUROC 0.897 on a 46.9% base rate, and for an intelligible reason: masked incidents withhold the channel identifying the cause, the agents reason from a correlate of a sibling family, and they diverge because the evidence genuinely underdetermines the answer. That divergence is the ensemble registering that the question is open, not a coordination failure.

On the second, disagreement carries essentially nothing, for the reason set out in Section IV: the harmful action was reversible, probe-carrying and unalarming, so the agents agreed it was low risk and were wrong together. Being wrong together produces low disagreement and high harm, which is the correlation observed.

This has a general form worth stating, because it is the paper’s main intellectual result. Disagreement among agents is an *epistemic* signal: it tracks whether the evidence determines an answer. Action risk in an operational setting is frequently not epistemic but *counterfactual*: it depends on a fact about the world, here whether a deployment was causal, that is absent from the evidence and about which the agents have no basis for disagreeing. Where the dangerous action is the one that looks reasonable to everybody, no amount of inter-agent disagreement will flag it. Architectures hoping to get safety out of disagreement are relying on danger being visible, and in the most important failure mode it is not.

B. Comparison with existing coordination strategies

The baselines separate by disposition rather than sophistication, occupying different points on one trade-off, and arbitration buys a better point on it rather than an escape from it: DARE asks for help on 6.2% of incidents against the 44.8% to 72.4% the cautious systems require, and against majority voting over the identical five answers the human cost differs by roughly a factor of seven for statistically indistinguishable safety, recovery and utility. For an operations team that is the difference between an assistant and a pager. Two baselines behaved notably: debate, the strongest coordination method tested, was among the least safe (14.06% harm) because mutual critique moved agents toward agreement, often on a confident wrong answer, the failure mode [18] and [19] describe from different directions. Self-consistency was indistinguishable from a single agent at five times the cost.

C. Safety versus autonomy

The frontier of Section IV is the honest form of the question, and it is steep where it matters: zero harm is achievable, at a coverage that resolves almost nothing. This is why we state a utility function rather than a safety rate alone, and vary the harm weight in Section VI. Reporting only unsafe-action rate would rank the confidence threshold first and a system that never acts above it; reporting only recovery would rank the single agent first while it harms nearly a fifth of incidents.

D. The role of incident memory

Incident memory is operationally useful and statistically unresolved, and the two should not be confused. Its fitted coefficient is negative (-0.547), which would mean corroborating history lowers the estimated harm score, but the bootstrap interval spans zero ([-1.424, 0.362], sign retained in 91% of 2,000 resamples), so this experiment does not identify the direction of that term. Removing memory lowers point-estimate utility (0.140 against 0.176), but that difference is not separated from zero either, and it cannot be ranked against removing disagreement, whose point estimate is effectively the same (0.135). What the evidence does support is informational rather than causal: retrieval was accurate enough to be useful and inaccurate enough to be realistic (74.5% top-1 family accuracy), and the memory agent is the component whose disagreement with the diagnosis agent is most informative, because diagnosis sees only present evidence while memory sees causes confirmed after the fact on incidents that once looked the same.

E. Corroboration against the public incident record

[t]

Fault family	Postmortems	Share (%)	Testbed (%)
schema change	4	2.0	12.5
resource exhaustion	35	17.6	12.5
config error	51	25.6	12.5
dependency failure	44	22.1	12.5
data quality	15	7.5	12.5
transformation logic	25	12.6	12.5
source delay	25	12.6	12.5
compound	0	0.0	12.5

Every result above comes from a simulator, including the oracle that decides what counts as harm, and that is this study’s largest threat to validity. Two of its claims can be checked against the public record without new inference. We coded 250 published postmortems, from infrastructure and SaaS providers including Amazon, Cloudflare, GitHub, CircleCI and CrowdStrike, against the fault taxonomy of Section III-H; 140 map to at least one family. The coding is released; the source narratives are cited rather than redistributed.

The dominant harm mode is not an artifact of the generator.

Section IV reports that 10 of 11 harmful executions were a deployment rollback on an incident no deployment caused, raising the obvious worry that a single generator choice produced the paper’s central negative result. The public record says otherwise: 11 of the 250 postmortems document a remediation that worsened or failed, at 8 distinct organisations, and several name rollback specifically: a CircleCI incident in which rolling back a type-changing deploy left the rows written between the two deploys unreadable, a GitHub report in which the rollback itself failed, a TravisCI report stating that the manual rollback broke. These are the structure the simulator produces: an action that is reversible, carries a probe, and looks reasonable to everyone, applied to an incident it does not fit.

The fault mix does not.

Table III is less comfortable. The testbed is uniform by construction, 12.5% per family, while the public record is skewed: config error accounts for 25.6% of labels and schema change for 2.0%. A uniform testbed over-samples rare families and under-samples common ones, so the aggregate rates here should not be read as estimates of what a deployed system would encounter. Two limits bound how far this reaches. The corpus is drawn from infrastructure and SaaS providers rather than data platforms, so the observed mix reflects the corpus’s domain as much as the world’s; and a postmortem records the action taken, never the outcomes of the actions that were not, so no public corpus supplies the counterfactual the safety analysis needs, a gap the agentic-operations benchmarks acknowledge in

their own terms [45], [46]. The corroboration here is of the mechanism and the distribution; the safety numbers themselves remain synthetic, and Section VII does not retreat from that.

F. Same-model versus heterogeneous agents

The sharpest evidence on whether role specialisation yields independent evidence is the self-consistency comparison in Table I: one model sampled five times out-pointed five differently-prompted roles on unsafe actions (0.572 against 0.422), a margin that does not reach significance on the independent sample ($p = 0.0896$). The defensible reading is not that sampling is better but that role prompting is not measurably better, which is already awkward for an architecture whose cost is the extra roles.

The heterogeneous arm (the extended results) sharpens this. Giving each role a *different* model produced more visible disagreement (1.69 distinct diagnoses per incident against 1.34), but the errors became *more* correlated, all three voices wrong together on 21.9% of incidents against 9.4%. Diversity bought disagreement without buying independence, and only the second makes a spread meaningful. Across all three configurations the diagnosis and remediation roles' errors stay strongly correlated whichever models fill them (0.731, 0.724, 0.864); incident memory is the role that decorrelates. The arm covers 32 incidents with smaller models and recovered 6.2% against 46.9%, so it cannot separate diversity from capability.

The roles are not redundant either: they disagree often, and that disagreement carries strong information about misdiagnosis. The conclusion is narrower than the architecture's premise. Role specialisation produces genuinely different *judgements about the cause*; it does not produce independent *estimates of danger*, because all five reason from the same evidence about the same absent fact.

Which other domains meet DARE's requirements is an analysis rather than a result, and is set out in the artifact.

6. Ablation and Sensitivity Analysis

Each ablation refits the risk model and retunes the thresholds on calibration under that ablation, then replays the *stored* agent outputs on the test split. No variant issues a model call, so no difference is attributable to sampling. Refitting rather than reusing the full coefficients matters: removing a signal changes what the remaining ones should weigh.

The replay evaluates a *single* arbitration round, since the gathering and replanning rounds were paid for by the full system and cannot be replayed for a variant that would have decided differently. The reference row is therefore single-round DARE (4.17% unsafe, 34.9% recovery, utility 0.176), not the headline configuration (5.73%, 39.1%, 0.149); comparisons within Table IV are like for like.

[t]

Variant	Unsafe	Recovery	Autonomy	Utility	Δ utility [95% CI]	p
	(%)	(%)	(%)	(mean)	vs full	
DARE (full)	4.17	34.9	34.9	0.176	—	—
— disagreement ($D = 0$)	4.17	31.2	31.2	0.135	-0.041 [-0.119, 0.038]	1
— incident memory	4.17	31.2	31.2	0.140	-0.036 [-0.115, 0.042]	1
— reversibility	4.17	34.9	34.9	0.166	-0.010 [-0.021, -0.003]	1

(no substitution)						
— blast radius ($B = 0$)	3.65	28.6	28.6	0.135	-0.041 [-0.137, 0.064]	1
— verification	5.73	39.6	39.6	0.147	-0.029 [-0.158, 0.088]	0.375
— confidence ($U = 0$, unweighted votes)	4.17	34.9	34.9	0.178	0.002 [-0.074, 0.078]	1

Before ranking anything, the limit of what these six comparisons can resolve. Each ablation’s utility difference from the full policy is a paired bootstrap over incidents, and 1 of 6 intervals excludes zero: the reversibility-based substitution step, which is also the smallest effect (-0.010). Removing disagreement and removing blast radius produce the same point estimate to three decimals (-0.041 and -0.041), and their intervals overlap almost entirely. Most ablations lower point-estimate utility, but only 1 of 6 differences is statistically separated from zero. The ablations therefore provide directional evidence; they do not establish that each term is individually necessary, and they do not support a ranking of signal importance. What follows is read as direction, not as attribution.

Blast radius and verification move the policy furthest.

Removing measured blast radius costs the most recovery of any ablation, 34.9% to 28.6%: the term is what lets the arbiter execute confidently where little downstream depends on the output. Removing verification moves the other way, raising unsafe from 4.17% to 5.73%, the highest of any variant, while raising recovery to 39.6%. Both are signals the *environment* supplies.

Disagreement: directionally useful, not statistically established.

Zeroing D lowers point-estimate utility, 0.176 to 0.135, and recovery, 34.9% to 31.2%, but the bootstrap interval on that utility difference includes zero. That is directionally consistent with a conditional contribution, and would not contradict RQ1 if real: a term can carry information given blast radius and history while carrying almost none marginally. With 18 harmful calibration events behind a five-coefficient fit, a signal of this size cannot be separated from overfitting, and we do not claim it.

Stated confidence contributes nothing.

Removing U and reverting to unweighted votes left the system marginally *better* (0.178 against 0.176). With the small fitted coefficient (0.364) this is a clean negative result: self-reported confidence added nothing to a procedure that already had disagreement, blast radius and history, though the same quantity was the *best single predictor* of harmful action in Table I. It is informative in isolation and redundant in combination.

Removing incident memory costs about as much as removing disagreement (0.140), and removing the reversibility-based substitution step costs a little (0.166) while leaving safety untouched, since substitution fired on only 6 incidents. That last is the one difference the bootstrap separates from zero, and it is the smallest of the six, which is a fair summary of how much this experiment can say about signal attribution.

A. Is the fitted risk model identified?

[t]

	Signal	Fitted	Bootstrap 95% CI	Sign	Identified
--	--------	--------	------------------	------	------------

α	D (combined disagreement)	1.179	[0.562, 1.788]	100%	yes
β	B (blast radius)	0.844	[0.081, 1.494]	98%	yes
γ	U (stated uncertainty)	0.364	[-0.049, 0.756]	96%	no
δ	$1 - V$ (unverifiability)	-0.944	[-1.579, -0.220]	100%	yes
ε	H (historical support)	-0.547	[-1.424, 0.362]	91%	no
<i>Does the harm score behave as a probability? Brier, base-rate Brier, calibration slope and intercept:</i>					
Calibration split (in-sample)		0.0626	base rate 0.0654	slope 2.26	int. 3.05
Test split (held out)		0.0623	base rate 0.0631	slope 1.31	int. 0.68

The risk model estimates five coefficients from 18 harmful outcomes, about 3.6 events per variable against a conventional floor of ten. At that ratio a logistic fit can return confident-looking coefficients that mostly record which events happened to land in the sample, so Table V asks what survives resampling. Two points of procedure first. The primary model is scikit-learn’s logistic regression at its default penalty, which is L_2 at $C = 1$; calling it simply logistic regression would imply an unpenalised fit that it is not. And the 2,000 bootstrap resamples are drawn over calibration incidents rather than over rows, because an incident contributes several action rows that share its incident-level signals.

The result qualifies the RQ5 coefficient story. Combined disagreement, blast radius and unverifiability keep their signs in essentially every resample and their intervals exclude zero. Stated uncertainty and historical support do not: their intervals are [-0.049, 0.756] and [-1.424, 0.362], and the history term keeps its fitted sign in 1813 of 2,000 resamples (91%). Two of five coefficients are therefore not identified by this calibration set, and the manuscript should not be read as having measured their direction. This is consistent with the ablation above, where removing stated confidence cost nothing.

The model is also scored as the probability it would need to be for the name to fit. On the held-out test split its Brier score is 0.0623 against 0.0631 for simply predicting the base rate, an improvement of about one part in eighty. Mean predicted harm is 7.3% against 6.8% observed, so the overall level is about right, but the calibration slope of 1.31 records real spread that the predictions do not capture. $R(a)$ is usable as a monotone score for ranking and thresholding, which is all the arbiter asks of it. It is not a calibrated probability, which is why Section III-E calls it an estimated harm score, and Section VII repeats the limitation.

Firth’s penalised likelihood is the textbook treatment at this event count and was not run: it lies outside the pinned environment, and a hand-written solver would be a worse risk in a paper whose argument is that its numbers can be checked. The omission is a limitation, not a result.

B. Do the fitted weights transfer across fault families?

Both splits are stratified by family, so the model can be refitted on seven families and scored on the eighth, rotated over all eight. The honest headline is that the analysis is underpowered rather than favourable: the 192 test incidents carry harmful outcomes at only a 6.8% base rate, two families contain none at all and three more rest on a single event, and over the 6 families where discrimination is computable the held-out fit averages 0.529 against 0.599 for the full model, both close to chance. What the fold structure does show is instability: withholding one family moves a coefficient by as much as 0.415 in absolute terms. Taken with the bootstrap above, this calibration set constrains the direction of three signals and the magnitude of none of them well. We report the analysis, reported in full in the extended results, because it was run and pre-specified, not because it supports the method.

C. Sensitivity

Three sensitivity analyses are summarised here and reported in full in the extended results. *Verifiability*: the most frequently proposed action, handing the incident to the upstream owner, carries no automated probe and is also never harmful; refitting with that action treated as verifiable moves the coefficient on $1 - V$ from -0.944 to -0.503, which locates the confound without removing it; the taxonomy itself is unchanged, since adjusting the environment until a coefficient takes the expected sign would not be an analysis. *Harm weight*: retuning across weights from -1 to -20 leaves the operating point unchanged from -1 through -5 at 5.73% harm and 39.1% autonomous recovery, then moves sharply to 0.52% at 14.6%, with the declared utility falling from 0.386 to -0.012 and crossing zero between. At the declared weight the sweep reproduces the headline operating point exactly. *Model scale*: repeated on claude-sonnet-4-5 over all 192 test incidents, DARE’s human-cost advantage widens (6.2% against 53.1%) while its utility advantage disappears, majority voting overtaking it (0.228 against 0.209). The human-cost result is the durable one.

7. Threats to Validity and Limitations

A. Construct validity

The environment is simulated.

Every safety claim here is a claim about outcomes assigned by a hand-written rule set over a hand-written topology. That is a real limitation and also what makes the study possible: on live infrastructure the counterfactual (what the declined action would have done) is unobservable, and without it abstention cannot be scored. The oracle is our model of consequence rather than a measurement of it, and we do not claim the absolute rates transfer to production. The design supports the *relative* comparison: all seven systems face identical incidents under an identical oracle.

Verifiability is confounded with action risk.

The fitted coefficient on $1 - V$ has the opposite sign to the one the risk model’s design anticipated, because of a property of our action space rather than a fact about verification: the most frequently proposed action carries no probe and is never harmful, while several probe-carrying actions are the ones the harm rules fire on. Section VI quantifies this by refitting with the handoff treated as verifiable. A cleaner separation would need an action space in which verifiability varies independently of consequence, which we did not design for.

Memory retrieval quality is a property of the simulator.

The store is generated by the process that generates the test incidents, so retrieval finds structurally similar cases more reliably than retrieval over real postmortems would. We reduced this by giving 14% of records a cause that later proved wrong, bringing top-1 family accuracy to 74.5% (81.7% direct, 65.9% masked), a plausible level, but one we chose rather than observed.

B. Internal validity

Replicates are not seeds, and there are fewer than planned.

The endpoint accepts no random seed, so a replicate is an independent sample rather than a reproducible draw; re-running the *analysis* reproduces our numbers exactly because every response is cached, but re-running the model calls would not. The study ran to a fixed inference budget, so the primary results rest on one full pass rather than the 2 planned, with a second pass on a reproducible 64-incident subsample estimating decoding stability. Every statistic assuming independent observations is computed on the 192 first-pass incidents, one decision each: pooling both passes would enter 64 incidents twice and narrow every interval by asking the same question again rather than by observing anything new. That is a real loss of precision compounding the low base rate below.

The harm base rate is low, and that bounds the power.

The agents propose an action the oracle marks harmful on roughly one incident in fourteen; frontier models are already conservative in action selection, and a low base rate is the honest consequence. RQ2 therefore rests on a modest number of discordant pairs with correspondingly wide intervals. We considered enriching the test set with adversarial states and did not, because manufacturing harm to obtain power would make the headline a property of our sampling. Where a comparison is underpowered we say so.

Calibration fitting is a design choice with consequences.

Fitting on the primary action alone produced too few harmful cases to identify five coefficients. Expanding to every proposed (incident, action) pair is defensible since the model scores pairs, but it shares incident-level signals across an incident’s rows, so the effective sample is smaller than the row count. Both fits are reported.

Ablations replay rather than re-sample.

Variants re-arbitrate stored agent outputs, which removes sampling variation so a difference is attributable to the arbitration change. It also means a variant is not charged for the extra rounds the full system paid for, which flatters the variants on cost but not on safety or recovery.

C. *External validity*

One model family.

The full comparison is repeated on a larger model of the same family over all 192 test incidents (Section VI), and that replication is where a substantive qualification appears: DARE’s advantage in human cost widens while its advantage in risk-adjusted utility disappears, majority voting overtaking it. A separate arm runs the role agents on local open-weight models of different families, but it covers 32 incidents with substantially smaller models, so “heterogeneous” is confounded with “weaker”. We have not established that any finding holds across providers.

The fitted risk model is weakly identified.

Five coefficients are estimated from 18 harmful outcomes, 3.6 events per variable. Under resampling (Section VI-A) three hold their sign and two (stated uncertainty and historical support) have intervals covering zero; withholding a fault family moves a coefficient by up to 0.415. Those two signs should not be read as measured. $R(a)$ is further not a calibrated probability, which is why it is called an estimated harm score: on held-out data its Brier score of 0.0623 barely improves on the base rate 0.0631, and its calibration slope is 1.31. It functions as a monotone score for ranking and thresholding, which is what the arbiter uses it for, and no claim beyond that is supported. Firth’s penalised likelihood, the standard treatment at this event count, was not run because it falls outside the pinned environment.

One domain, and one within-testbed holdout.

The experiments are entirely within enterprise data-platform recovery. The leave-one-fault-family-out analysis (Section VI-B) is the only transfer evidence and it is weak: it stays inside the same simulator, two of eight families contain no harmful outcome, and the families where discrimination is computable rest on one to four events. It tests whether the fitted weights are family-specific; it says nothing about another platform, organisation or a real incident stream, and it does not reduce the synthetic-testbed limitation above. The generalization analysis in the artifact states the framework’s requirements so a reader can judge which domains satisfy them, and is explicitly speculative. No experiment here tests any other domain.

Prompts are a confound we did not isolate.

Role behaviour depends on role prompts, which we wrote; a different phrasing of the risk agent’s brief would change its scores and therefore D_{risk} . We ran no prompt-sensitivity study. The prompts are published verbatim in the artifact so the sensitivity can be measured by others.

D. *Infrastructure and reproducibility*

The endpoint used was a third-party proxy that exhausted its quota mid-study and returned a streamed response to a non-streaming request, both handled in the client. Neither affects the results, since every response is cached, but wall-clock latency includes queueing at a shared endpoint and should be read as an upper bound rather than a property of the method; token and call counts are endpoint-independent and are the better cost measure. Provider-side prompt caching never engaged, because this model tier requires a 2,048-token cacheable prefix while the role prompts are approximately 1,800, so every call was billed at the uncached rate, the direct cause of the reduced replication above.

8. Conclusion

On 192 enterprise data-platform incidents with root cause, action consequence and blast radius supplied by a deterministic simulator rather than a model, the answer separates in two. Disagreement is a strong signal of uncertainty about the *cause*: diagnostic disagreement discriminates misdiagnosis at AUROC 0.897, because the incidents on which agents diverge are those whose evidence genuinely underdetermines the answer, so treating that divergence as noise to be resolved by convergence discards real information.

Disagreement is not a usable signal of *action risk*. No channel separated the incidents where the proposed action would cause harm in the direction the hypothesis predicted, and risk disagreement was significantly *inversely* related to harm. That inverse relationship is evidence about the mechanism, not a safety detector in disguise: a system acting on it would execute most readily where the agents dispute the danger most. It arises because the harm came almost entirely from a reversible, verifiable action every agent agreed was safe and that was wrong for a reason absent from the evidence. We take this to generalise: where the dangerous action is the one that looks reasonable, inter-agent dispute cannot detect it, and an architecture expecting safety to fall out of disagreement is relying on danger being visible.

Disagreement-aware arbitration nevertheless produced the best safety/autonomy trade-off of the seven systems under the primary model and the declared utility weights, matching majority voting on safety and recovery while requiring human intervention on 6.2% of incidents rather than 44.8%. That ranking is objective- and model-dependent: on the larger-model arm majority voting overtakes DARE on utility, while the human-cost gap widens in DARE's favour. The ablations give directional evidence only, and the fitted risk model is weakly identified at this event count (Sections VI and VII).

For practice this reframes where the effort belongs (the generalization analysis in the artifact): measuring disagreement is nearly free and worth doing, but for triage and evidence gathering rather than as a safety interlock. An autonomous recovery system that wants to act needs consequence information the agents cannot supply about themselves. The limits are stated in Section VII. The work the results point to is not a better disagreement measure but a better *counterfactual* one: an estimate of whether a recent change was causal, a question about the world rather than about the ensemble. The artifact, including every cached model response, is released so that question can be pursued against the same incidents.

9. Data and Code Availability

All data are synthetic and generated by the released code. The implementation, agent prompts, cached model responses, statistical analysis and results files are released under a permissive licence at *[repository URL withheld for blind review]*. Re-running the analysis reproduces every number in this paper offline, without issuing a model call.

10. Generative AI Use Disclosure

Language models are the object of study: Claude Haiku 4.5 as the agent model, Claude Sonnet 4.5 for the model-scale comparison, and local open-weight models for the heterogeneous arm. Exact identifiers, temperatures and endpoints are in the released results files. Generative AI tools were additionally used in a general supporting capacity during preparation, and not to generate, estimate, impute or adjust any experimental result: every number is produced by the released code from the released data, and no measured value is typed into the prose. No AI system is an author.

11. Conflict of Interest

The author declares no financial or non-financial conflict of interest. The research was conducted independently of the author’s employment, uses no proprietary, confidential or customer data, and received no funding.

1. S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. H. Dodds, N. DasSarma, E. Tran-Johnson, *et al.*, “Language models (mostly) know what they know.” 2022. Available: <https://arxiv.org/abs/2207.05221>
2. S. C. Lin, J. Hilton, and O. Evans, “Teaching models to express their uncertainty in words,” *Trans. Mach. Learn. Res.*, 2022. Available: <https://arxiv.org/abs/2205.14334>
3. M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi, “Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs,” in *Proceedings of the International Conference on Learning Representations*, 2024. Available: <https://arxiv.org/abs/2306.13063>
4. Y. Ruan, H. Dong, A. Wang, S. Pitis, Y. Zhou, J. Ba, Y. Dubois, C. J. Maddison, and T. Hashimoto, “Identifying the risks of LM agents with an LM-Emulated sandbox,” in *Proceedings of the International Conference on Learning Representations*, 2024. Available: <https://arxiv.org/abs/2309.15817>
5. Z. Zhang, S. Cui, Y. Lu, J. Zhou, J. Yang, H. Wang, and M. Huang, “Agent-SafetyBench: Evaluating the safety of LLM agents.” 2024. Available: <https://arxiv.org/abs/2412.14470>
6. Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, *et al.*, “AutoGen: Enabling Next-Gen LLM applications via Multi-Agent conversation.” 2023. Available: <https://arxiv.org/abs/2308.08155>
7. S. Hong, M. Zhuze, J. Chen, X. Zheng, Y. Cheng, C. Zhang, J. Wang, Z. Wang, S. K. S. Yau, Z. Lin, *et al.*, “MetaGPT: Meta programming for a Multi-Agent collaborative framework.” 2023. Available: <https://arxiv.org/abs/2308.00352>
8. T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, “Large language model based Multi-Agents: A survey of progress and challenges.” 2024. Available: <https://arxiv.org/abs/2402.01680>
9. Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, “Improving factuality and reasoning in language models through multiagent debate,” in *Proceedings of the International Conference on Machine Learning*, 2024. Available: <https://arxiv.org/abs/2305.14325>
10. T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, and Z. Tu, “Encouraging divergent thinking in large language models through Multi-Agent debate,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2024, pp. 17889–17904. <https://doi.org/10.18653/v1/2024.emnlp-main.992>
11. X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H., S. Narang, A. Chowdhery, and D. Zhou, “Self-Consistency improves chain of thought reasoning in language models,” in *Proceedings of the International Conference on Learning Representations*, 2023. Available: <https://arxiv.org/abs/2203.11171>
12. M. K. Sivaraman, “The incident memory layer: An architectural component for lakehouse reliability,” *Journal of Information Systems Engineering and Management*, vol. 9, no. 4, 2024.
13. M. K. Sivaraman, “The reliability gap in modern lakehouse architectures: A failure fingerprinting framework for operational intelligence,” *Current Journal of Applied Science and Technology*, vol. 45, no. 3, pp. 36–55, 2026. <https://doi.org/10.9734/cjast/2026/v45i34673>
14. M. K. Sivaraman, “Closed-loop LLM-driven self-healing for enterprise data pipelines: Architecture and preliminary evaluation,” in *Proceedings of the IEEE Technology and Innovation Conference (TIC)*, 2026.
15. S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, “ReAct: Synergizing reasoning and acting in language models,” in *Proceedings of the International Conference on Learning Representations*, 2023. Available: <https://arxiv.org/abs/2210.03629>
16. N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: Language agents with verbal reinforcement learning,” in *Advances in Neural Information Processing Systems 36*, Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023, pp. 8634–8652. <https://doi.org/10.52202/075280-0377>
17. G. Irving, P. F. Christiano, and D. Amodei, “AI safety via debate.” 2018. Available: <https://arxiv.org/abs/1805.00899>
18. A. Smit, P. Duckworth, N. Grinsztajn, T. D. Barrett, and A. Pretorius, “Should we be going MAD? A look at Multi-Agent debate strategies for LLMs.” 2023. Available: <https://arxiv.org/abs/2311.17371>
19. M. Cemri, M. Z. Pan, S. Yang, L. A. Agrawal, B. Chopra, R. Tiwari, K. Keutzer, A. Parameswaran, D. Klein, K. Ramchandran, *et al.*, “Why do Multi-Agent LLM systems fail?” in *Advances in Neural Information Processing Systems 38*, Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025, pp. 135676–135729. <https://doi.org/10.52202/085713-4082>
20. L. Yoffe, A. Amayuelas, and W. Y. Wang, “DebUnc: Improving large language model agent communication with uncertainty metrics,” in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Association for Computational Linguistics, 2025, pp. 23299–23315. <https://doi.org/10.18653/v1/2025.findings-emnlp.1265>
21. B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles.” 2016. Available: <https://arxiv.org/abs/1612.01474>
22. Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *Proceedings of the International Conference on Machine Learning*, 2016. Available: <https://arxiv.org/abs/1506.02142>
23. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proceedings of the International Conference on Machine Learning*, 2017. Available: <https://arxiv.org/abs/1706.04599>

24. L. Kuhn, Y. Gal, and S. Farquhar, "Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation," in *Proceedings of the International Conference on Learning Representations*, 2023. Available: <https://arxiv.org/abs/2302.09664>
25. C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, 1970. <https://doi.org/10.1109/tit.1970.1054406>
26. R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," *Journal of Machine Learning Research*, 2010.
27. Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks." 2017. Available: <https://arxiv.org/abs/1705.08500>
28. D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and Out-of-Distribution examples in neural networks," in *Proceedings of the International Conference on Learning Representations*, 2017. Available: <https://arxiv.org/abs/1610.02136>
29. B. Wen, J. Yao, S. Feng, C. Xu, Y. Tsvetkov, B. Howe, and L. L. Wang, "Know your limits: A survey of abstention in large language models," *Transactions of the Association for Computational Linguistics*, vol. 13, pp. 529–556, 2025. https://doi.org/10.1162/tacl_a_00754
30. Y. A. Yadkori, I. Kuzborskij, D. Stutz, A. György, A. Fisch, A. Doucet, I. Beloshapka, W.-H. Weng, Y.-Y. Yang, C. Szepesvári, *et al.*, "Mitigating LLM hallucinations via conformal abstention." 2024. Available: <https://arxiv.org/abs/2405.01563>
31. S. Feng, W. Shi, Y. Wang, W. Ding, V. Balachandran, and Y. Tsvetkov, "Don't hallucinate, abstain: Identifying LLM knowledge gaps via Multi-LLM collaboration," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2024, pp. 14664–14690. <https://doi.org/10.18653/v1/2024.acl-long.786>
32. D. Madras, T. Pitassi, and R. S. Zemel, "Predict responsibly: Improving fairness and accuracy by learning to defer," in *Proceedings of the Neural Information Processing Systems*, 2018. Available: <https://arxiv.org/abs/1711.06664>
33. H. Mozannar and D. Sontag, "Consistent estimators for learning to defer to an expert." 2020. Available: <https://arxiv.org/abs/2006.01862>
34. S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, *et al.*, "Guidelines for Human-AI interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM, 2019, pp. 1–13. <https://doi.org/10.1145/3290605.3300233>
35. A. Z. Ren, A. Dixit, A. A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley, *et al.*, "Robots that ask for help: Uncertainty alignment for large language model planners," in *Proceedings of the Conference on Robot Learning*, 2023. Available: <https://arxiv.org/abs/2307.01928>
36. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety." 2016. Available: <https://arxiv.org/abs/1606.06565>
37. S. Nailin, D. Atkinson, M. Green, M. Hamadi, C. Swift, D. Schonholtz, A. T. Kalai, and D. Bau, "Testing language model agents safely in the wild." 2023. Available: <https://arxiv.org/abs/2311.10538>
38. J. O. Kephart and D. M. Chess, "The vision of autonomic computing," *Computer*, vol. 36, no. 1, pp. 41–50, 2003. <https://doi.org/10.1109/mc.2003.1160055>
39. P. Notaro, J. Cardoso, and M. Gerndt, "A survey of AIOps methods for failure management," *ACM Transactions on Intelligent Systems and Technology*, vol. 12, no. 6, pp. 1–45, 2021. <https://doi.org/10.1145/3483424>
40. T. Ahmed, S. Ghosh, C. Bansal, T. Zimmermann, X. Zhang, and S. Rajmohan, "Recommending Root-Cause and mitigation steps for cloud incidents using large language models," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, IEEE, 2023, pp. 1737–1749. <https://doi.org/10.1109/icse48619.2023.00149>
41. Y. Chen, H. Xie, M. Ma, Y. Kang, X. Gao, L. Shi, Y. Cao, X. Gao, H. Fan, M. Wen, *et al.*, "Automatic root cause analysis via large language models for cloud incidents," in *Proceedings of the Nineteenth European Conference on Computer Systems*, ACM, 2024, pp. 674–688. <https://doi.org/10.1145/3627703.3629553>
42. Z. Wang, Z. Liu, Y. Zhang, A. Zhong, J. Wang, F. Yin, L. Fan, L. Wu, and Q. Wen, "RCAgent: Cloud root cause analysis by autonomous agents with Tool-Augmented large language models," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, ACM, 2024, pp. 4966–4974. <https://doi.org/10.1145/3627673.3680016>
43. D. Roy, X. Zhang, R. Bhav, C. Bansal, P. Las-Casas, R. Fonseca, and S. Rajmohan, "Exploring LLM-Based agents for root cause analysis," in *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, ACM, 2024, pp. 208–219. <https://doi.org/10.1145/3663529.3663841>
44. X. Zhou, G. Li, Z. Sun, Z. Liu, W. Chen, J. Wu, J. Liu, R. Feng, and G. Zeng, "D-Bot: Database diagnosis system using large language models," *Proceedings of the VLDB Endowment*, vol. 17, no. 10, pp. 2514–2527, 2024. <https://doi.org/10.14778/3675034.3675043>
45. Y. Chen, M. Shetty, G. Somashekar, M.-J. Ma, Y. L. Simmhan, J. Mace, C. Bansal, R. Wang, and S. Rajmohan, "AIOpsLab: A holistic framework to evaluate AI agents for enabling autonomous clouds," *Conference on Machine Learning and Systems*, 2025. Available: <https://arxiv.org/abs/2501.06706>
46. S. Jha, R. Arora, Y. Watanabe, T. Yanagawa, Y. Chen, J. Clark, B. Bhavya, M. Verma, H. Kumar, H. Kitahara, *et al.*, "ITBench: Evaluating AI agents across diverse Real-World IT automation tasks," in *Proceedings of the International Conference on Machine Learning*, 2025. Available: <https://arxiv.org/abs/2502.05352>

47. M. Zaharia, A. Ghodsi, R. Xin, and M. Armbrust, "Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics," in *Conference on Innovative Data Systems Research*, 2021.
48. M. Armbrust, T. Das, L. Sun, B. Yavuz, S. Zhu, M. Murthy, J. Torres, H. van Hovell, A. Ionescu, A. Łuszczak, *et al.*, "Delta lake: High-Performance ACID table storage over cloud object stores," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 3411–3424, 2020. <https://doi.org/10.14778/3415478.3415560>
49. S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Graffberger, "Automating large-scale data quality verification," *Proceedings of the VLDB Endowment*, vol. 11, no. 12, pp. 1781–1794, 2018. <https://doi.org/10.14778/3229863.3229867>
50. E. Breck, M. Zinkevich, N. Polyzotis, S. E. Whang, and S. Roy, "Data validation for machine learning," in *Proceedings of Machine Learning and Systems (MLSys)*, 2019, pp. 334–347. Available: <https://openalex.org/W2946595616>
51. C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
52. K. Krippendorff, *Content analysis: An introduction to its methodology*. SAGE Publications, 2019. <https://doi.org/10.4135/9781071878781>
53. Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947. <https://doi.org/10.1007/bf02295996>
54. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, p. 837, 1988. <https://doi.org/10.2307/2531595>
55. S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 1979. <https://doi.org/10.2307/4615733>