Shared Corruption Identity and Cross-Route Exact-Repair Failure Dependence

A Finite Synthetic Eight-Route Study

Nicholas Zinner Beacon Bot

Future Shock (future-shock.ai)

September 09, 2026

Public research working paper. Not peer reviewed. Findings apply to the finite synthetic route panel described here.

Abstract

In this finite synthetic eight-route study, mean unsuccessful-exact-repair failure changed little, while the all-eight failure rate increased. We asked whether giving every language-model deployment route the same complete corruption variant increased dependence in observable non-detection, unsuccessful exact repair, or terminal task failure relative to independently and identically distributed (IID), with-replacement assignment from the same corruption pool. Here, a deployment route was the specific model/provider/deployment endpoint tested at the collection epoch. The matched design retained arithmetic, constraint, extraction, and tool-call tasks, with code excluded prospectively after calibration failure. Of 1,800 item families, the frozen primary analysis retained 1,677 with complete support across the clean, shared-corruption, and IID-corruption conditions; only the two corrupt arms entered the dependence contrast. Based on 9,999 stratified item-family bootstrap resamples, the shared-to-IID dependence-inflation ratio for unsuccessful exact repair was 1.129, with simultaneous interval [1.060, 1.203]. The unsuccessful-exact-repair interval was above one in the frozen implemented primary analysis. The study nevertheless did not satisfy its full confirmatory-success rule because the point estimator used equal item weights rather than the protocol's fixed equal stratum weights, and because a required precollection calibration-support projection gate was omitted and retrospectively fails. The primary intervals for non-detection, 1.038 [0.948, 1.137], and terminal failure, 1.061 [0.992, 1.135], both included one. In postconfirmatory analyses, mean unsuccessful-repair failure changed from 59.23% to 59.91%, with a mean-difference interval crossing zero, whereas all-eight failure changed from 4.59% to 6.98%. Named weighting, support, and missing-cell scenarios retained unsuccessful-repair failure-dependence ratios above one; an unrestricted binary-completion envelope of 0.989–1.311, however, included the null and is not a confidence interval. The result concerns dependence in this finite synthetic route panel, not a proportional change in marginal risk, incidents, losses, or an actuarial multiplier.

Keywords: model fleets; correlated failure; evidence corruption; exact repair; dependence inflation; missing data

1 Introduction

Using several models is often treated as insurance against any one model's errors, but the protection depends on whether the models bring information that is not duplicated in one another's

mistakes. Classical work makes this condition explicit: identical predictors add no information, whereas disagreement can reduce ensemble error when individual accuracy is held fixed (Krogh and Vedelsby, 1994; Dietterich, 2000). Model-fleet evaluations often emphasize route accuracy, average vote accuracy, or winner selection, all of which can overlook an all-member failure on the same item. Joint-tail events can therefore be informative, although this study did not measure their real-world frequency, severity, or deployment utility.

This study tests one possible source of concentrated failure: exposing every route to the same complete corruption variant. A deployment route is the specific model/provider/deployment endpoint tested at the collection epoch. We compare a panel in which all routes receive the same particular corruption with the same panel receiving IID, with-replacement draws from a common, severity-matched corruption pool. Both arms are corrupt, and every packet contains one corrupted field. They differ only in whether routes share a variant or draw variants independently. The design separates sharing structure from corruption dose and marginal variant-selection probability, but it cannot separate the attributes bundled within each variant.

Recent language-model methods combine samples, candidates, judges, or debating agents in different ways. Self-consistency aggregates reasoning paths (Wang et al., 2023); trained verifiers rank candidate solutions (Cobbe et al., 2021); process supervision evaluates intermediate steps (Lightman et al., 2023); LLM-Blender ranks and fuses outputs from several models (Jiang et al., 2023); and multiagent debate lets instances exchange proposed answers over several rounds (Du et al., 2024). These methods improved task performance in their tested settings, but those improvements do not establish independence among participating outputs or rule out failures aligned by shared evidence. An evaluation of more than 350 language models found substantial error correlation, including across distinct architectures and providers, and examined downstream effects in model-as-judge evaluation and a simulated hiring setting (Kim et al., 2025). Our controlled contrast is narrower: all-shared versus IID-with-replacement corruption-variant assignment in one frozen synthetic design.

The study estimates cross-route dependence for three observable failure outcomes under that contrast. In the frozen implemented primary analysis, the unsuccessful-exact-repair interval was above one, while the primary intervals for non-detection and terminal failure included one. The postconfirmatory summaries also moved differently: all-eight repair failure increased, mean repair failure changed little, and the at-least-half rate did not increase. Weighting, support, and missingness analyses qualify these findings rather than provide separate contributions. In particular, the implemented point-estimator weighting differs from the protocol text, and the omitted calibration-support projection gate retrospectively fails. The estimates remain defined, but the study did not meet the protocol's full confirmatory-success rule.

2 Related work

2.1 Diversity, dependence, and ensemble error

Ensemble learning has long distinguished individual predictive quality from dependence among errors. Krogh and Vedelsby expressed regression ensemble error as weighted average member error minus an ambiguity term, thereby making disagreement part of the performance calculation (Krogh and Vedelsby, 1994). Dietterich reviewed voting ensembles and several mechanisms through which collections of classifiers can outperform single systems (Dietterich, 2000). For classification, however, "diversity" has no single agreed measure. Kuncheva and Whitaker compared ten pairwise and non-pairwise statistics and found no straightforward relationship with ensemble accuracy in practical pattern-recognition problems (Kuncheva and Whitaker, 2003). No one dependence statistic

should therefore be read as a universal quality index.

In this paper, dependence inflation has a narrower role. It compares the observed item-to-item variance of the equal-weight route-panel failure fraction with the variance expected from route marginal variances alone. The shared-to-IID ratio then asks whether failures are more concentrated when routes share the same complete corruption variant. It neither ranks general ensemble diversity nor identifies a latent correlation mechanism or deployment loss function.

2.2 Language-model aggregation and verification

Language-model aggregation includes architectures with quite different dependence structures. Self-consistency samples several reasoning paths from one model and marginalizes over their answers (Wang et al., 2023). Candidate-verifier systems generate many completions and use a learned judge to select one (Cobbe et al., 2021), while process reward models evaluate intermediate reasoning steps rather than only the final answer (Lightman et al., 2023). Other systems rank and fuse outputs across models (Jiang et al., 2023) or allow instances to communicate before producing a final answer (Du et al., 2024). The architectures differ in model identity, prompt sharing, communication topology, evaluation target, and access to independent evidence.

Those differences shape what each result can support. A larger candidate set may contain a correct answer even when a common verifier ranks it incorrectly. Debate can expose disagreement, but similar capabilities or responses can also yield static dynamics and convergence on a shared misconception (Estornell and Liu, 2024). That finding concerns communicating debaters, whereas the routes here received stateless packets and did not communicate. Cross-model routing can use heterogeneous strengths, but provider and architecture labels remain imperfect proxies for empirical error diversity. Kim and colleagues measured correlated errors across a large contemporary model set rather than inferring independence from branding (Kim et al., 2025). A capability-controlled audit similarly found that several diversity statistics were entangled with capability and that a modest pairwise co-failure association remained after adjustment (Kim, 2026). Both studies are observational evaluations of model panels, voting, or judge reliability. By contrast, this experiment manipulates complete corruption-variant sharing across eight fixed routes and prespecifies dependence contrasts for non-detection, exact-repair failure, and terminal failure. It does not claim to discover correlated errors generally or a universal divergence between average and joint failure.

2.3 Misinformation-bearing context

Controlled and benchmark studies show that language models remain susceptible to misinformation-bearing or knowledge-conflicting context. MisBench, for example, varies factual, temporal, semantic, and stylistic properties of misinformation and evaluates model detection and answer behavior (Peng et al., 2025). Such studies establish that contextual conflict can alter individual-model behavior and motivate evaluation of defenses. They do not test all-shared versus IID-with-replacement assignment or cross-route dependence. In the present design, corruption dose and marginal variant-selection probability are fixed while sharing structure changes; the misinformation literature provides context for that treatment but does not validate this study's outcomes.

2.4 Missingness and partial identification

Infrastructure failures create missing model cells, which should not be recoded automatically as behavioral failures. Standard missing-data analysis distinguishes among mechanisms and shows that complete-case analysis, weighting, imputation, and sensitivity methods answer different questions under different assumptions (Little and Rubin, 2019). Under deliberately weak assumptions

about unobserved outcomes, partial-identification bounds can replace a preferred completion model (Manski, 1990). We preserve that distinction: named deterministic completions serve as specification checks, while the unrestricted binary-completion envelope ranges over possible missing binary outcomes. Because that envelope is not a confidence interval and includes one, it does not support robustness to arbitrary outcome-dependent missingness.

3 Study design and methods

3.1 Research question and experimental unit

We asked a finite question: within the frozen route panel and calibrated short-state population, does sharing one complete corruption variant increase dependence in non-detection, unsuccessful exact repair, or terminal failure relative to IID-with-replacement assignment from the same corruption pool? (CMFC D1 Protocol, 2026) Each base item paired a machine-checkable task and canonical state with a structured evidence packet containing several targetable fields. A treatment variant changed exactly one field, left the canonical answer unchanged, and preserved visible information from which the clean value could be recovered.

Three corruption mechanisms were retained. One placed a retrieved claim in conflict with a verified record, another placed a tool observation in conflict with replayable or redundant state, and the third placed stale state in conflict with a versioned update log. Following a second scientific calibration failure, Amendment 0021 changed qualification prospectively, before pilot or confirmation materialization (CMFC D1 Amendment 0021, 2026). Qualification was pooled by mechanism and predicted difficulty across the arithmetic, constraint, extraction, and tool-call domains. The final ranges were [0.30, 0.70] for clean terminal success, [0.20, 0.80] for corrupt terminal success, and [0.10, 0.90] for corrupt detection and repair success. Code was excluded as a whole construction tier. The planned retained population was balanced at 600 items per mechanism, and the calibration, pilot, and confirmation populations remained fresh and disjoint.

Figure 1 illustrates one matched item family with a synthetic example, not a recovered prompt, response, item, or model quote. A short task has a correct underlying state, while its structured packet contains one wrong field and visible evidence that allows correction. All eight routes receive the same complete wrong variant in the shared arm. In the IID arm, each route draws with replacement from the same pool, so accidental matches remain possible. Each response is scored separately for flagging the packet, exactly replacing the designated field, and returning the correct task answer. An *item family* comprises the matched clean, shared-corruption, and IID-corruption versions of one base task, with one response from each retained route in each condition.

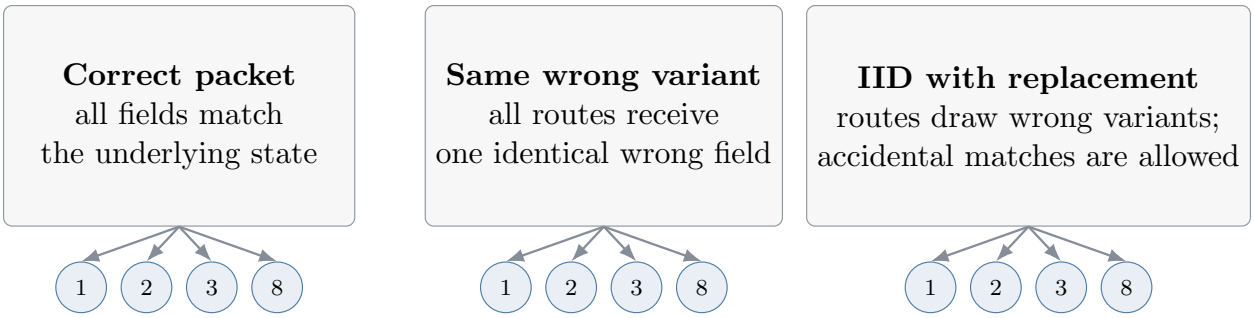

Figure 1: Synthetic explanatory schematic of one matched item family. The shared and IID arms differ in complete corruption-variant assignment, not corruption count, packet structure, or canonical answer. Route labels are schematic and do not reproduce an observed trial.

3.2 Conditions, routes, and support

The three conditions used structurally matched clean packets (E_{clean}), one complete corruption variant shared across all routes ($E_{corrupt\_shared}$), and routewise IID draws with replacement from a frozen severity-matched pool ($E_{corrupt\_iid}$). Every base item had at least twelve valid variants, and the selected shared variant had the same marginal selection probability as each IID draw. Thus the estimand compares all-shared with IID-with-replacement assignment, not shared evidence with guaranteed-distinct or independently sourced evidence. Collisions between IID draws were possible. Although the protocol designated duplicates as valid and required them to be reported, the bounded shareable artifacts do not contain the realized collision count or topology. Their effect on the direction or magnitude of the nonlinear dependence contrast is therefore unidentified. Assignment followed a frozen seed and was balanced by mechanism, domain, predicted difficulty, and target position. Prompts were stateless and text-only, with tools, search, fallback routing, conversational memory, and cross-cell carryover disabled (CMFC D1 Protocol, 2026).

The study reached its final eight-route panel through a documented sequence of infrastructure decisions and continuations. The original ten-route pilot failed an infrastructure gate, so one DeepInfra route was excluded in full after availability failures; a disjoint nine-route pilot then passed. The nine-route confirmation received a wall-time continuation that preserved accepted and terminal-missing cells. A draft continuation that changed 5,400 Z.AI request digests failed closed before event import or provider calls. Nvidia/CoreWeave was subsequently excluded in full and outcome-blind after condition-specific support ceilings fell below the required floor, and the retained eight-route run later received a second wall-time continuation. Each decision used delivery state and hash equality without outcome access (CMFC D1 Primary Report, 2026). Appendix A lists the retained routes verbatim. These identifiers denote the model/provider/deployment endpoints tested at the collection epoch, not timeless model families or evidence of provider independence. The design collected one behavioral generation per retained route-item-condition cell; infrastructure retries continued the same logical cell rather than creating repeated behavioral draws.

Of 1,800 planned item families, 1,677 had exact support for all eight routes in all three conditions. This all-arm requirement defined the primary analysis set, although the dependence ratio compares only the shared and IID corrupt arms; the clean arm is the matched baseline for separate corruption-versus-clean marginal summaries. The archive records 43,073 accepted cells out of 43,200 planned, or 99.71% overall support, and a minimum route-condition support of 99.39%

(CMFC D1 Primary Report, 2026). Infrastructure-missing cells were not scored as model failures. By contrast, an accepted provider response that violated the response contract could be scored as a behavioral failure. Transport, routing, authentication, ambiguous-delivery, scorer, and related infrastructure states remained separate.

3.3 Observable outcomes

The response contract required a single JSON object with a decision, target identifier, corrected value, repair object, and terminal answer. A fully correct response to a corrupt packet had to flag the packet, identify the corrupted target, supply its clean replacement, emit the exact repair object, and return the canonical answer. The three primary failures were:

- **Non-detection:** the corrupt packet was not correctly flagged under the observable decision contract.
- **Unsuccessful exact repair:** the designated replacement object was not emitted exactly.
- **Terminal task failure:** the final answer was incorrect, scored independently of earlier fields.

Exact repair is narrower than practical recovery in general. A route can return the canonical answer without emitting the designated repair object, just as a valid flag can accompany incorrect localization or replacement. The outcomes record emitted behavior, not latent awareness, belief, or a cognitive sequence.

3.4 Dependence estimand

For route m, item i, condition c, and a binary failure outcome, let $F_{m,i,c} = 1$ for failure and 0 for success. With $M = 8$ and equal route weights $w_m = 1/M$, dependence inflation is

$$DI_c = \frac{\mathrm{Var}_i\left(\sum_{m=1}^{M} w_m F_{m,i,c}\right)}{\sum_{m=1}^{M} w_m^2 \,\mathrm{Var}_i(F_{m,i,c})}. \tag{1}$$

The numerator is the item-to-item variance of the route-panel failure fraction. The denominator preserves route marginal variances while omitting cross-route covariance terms. A value above one therefore indicates net positive covariance in the tested matrix, whereas values below one can occur under net negative dependence. The primary contrast was

$$\theta = \log\left(\frac{DI_{shared}}{DI_{iid}}\right), \tag{2}$$

The ratio is reported after exponentiating this log contrast. Thus 1.129 means that the estimated dependence-inflation index is 12.9% higher in the shared arm, not that marginal failure, incident probability, loss, or capital requirements are 12.9% higher.

For the primary point estimates, the frozen implementation gave equal weight to each of the 1,677 joint-complete item families. The bootstrap remained paired across corrupt conditions and stratified by mechanism, retained domain, and frozen predicted difficulty. Each of the 9,999 resamples preserved the retained stratum counts. This procedure preserves observed composition during resampling; it does not impose the protocol's fixed equal weights over mechanisms, domains, and difficulty bands on the point estimator. Although the planned population was balanced at 600 items per retained mechanism after whole-domain code exclusion, complete-case deletion changed the realized stratum composition.

Consequently, 1.129 [1.060, 1.203] is the simultaneous interval from the frozen implemented equal-item analysis, not from a fully protocol-conforming equal-stratum analysis. The three implemented primary contrasts used simultaneous intervals based on a frozen 0.975 studentized max-absolute critical quantile (CMFC D1 Protocol, 2026; CMFC D1 Primary Report, 2026). The closest retained equal-stratum sensitivity gave equal mass to mechanism, domain, and retained difficulty tier. Because it was examined postconfirmatorily, it does not inherit the simultaneous familywise status of the primary analysis.

These ratios are finite-set descriptive contrasts for the retained routes and items. Their empirical item-family bootstrap intervals condition on the eight fixed routes, frozen construction and corruption pools, retained complete-case support, and observed stratum composition. They describe stability to resampling item families within those strata, not uncertainty over routes, providers, future model versions, or real-world task populations. Because the design-simulation bytes are unavailable, familywise coverage cannot be independently established beyond the retained design receipt.

3.5 Postconfirmatory analyses

The frozen publication closeout examined equal-mechanism and calibration-composition weighting, condition-specific support, six deterministic missing-cell completions, and a conservative unrestricted binary-completion envelope. It also compared mean endpoint failure with all-eight failure (CMFC D1 Publication Closeout, 2026). All intervals from these analyses are postconfirmatory percentile intervals rather than primary simultaneous intervals.

The corrected observable-stage analysis used 1,716 items complete across both corrupt arms and, separately, 1,758 clean items. It applied stage-specific event/non-event, variance, DI, bootstrap, and all-eight estimability gates, including a 20-failure/20-success endpoint rule (CMFC D1 Stage V2, 2026). The separate infrastructure V3 analysis examined time-matched endpoint pairs under multiple windows and co-optimal matchings using historical delivery records rather than model-behavior outcomes (CMFC Infrastructure V3, 2026). Timing, scheduling, routing, and retries were not randomized for the infrastructure question.

4 Results

4.1 Primary dependence family

Table 1 and Figure 2 present the frozen primary family. For unsuccessful exact repair, the shared-to-IID ratio was 1.129 with simultaneous interval [1.060, 1.203]. The simultaneous intervals for non-detection and terminal failure included one. The global familywise Monte Carlo value was $p = 0.0001$, the minimum attainable plus-one value with 9,999 resamples. Rejection of the family-level null does not imply that every component is positive (CMFC D1 Primary Report, 2026; CMFC D1 Publication Closeout, 2026).

Table 1: Frozen primary shared-to-IID dependence-inflation ratios. Simultaneous intervals use 9,999 frozen stratified item-family resamples. Population: 1,677 joint-exact-support item families and eight routes.

Outcome	DI_{shared}	DI_{iid}	Ratio [simultaneous interval]
Non-detection failure	1.684	1.623	1.038 [0.948, 1.137]
Unsuccessful exact repair	2.483	2.199	1.129 [1.060, 1.203]
Terminal task failure	2.030	1.913	1.061 [0.992, 1.135]

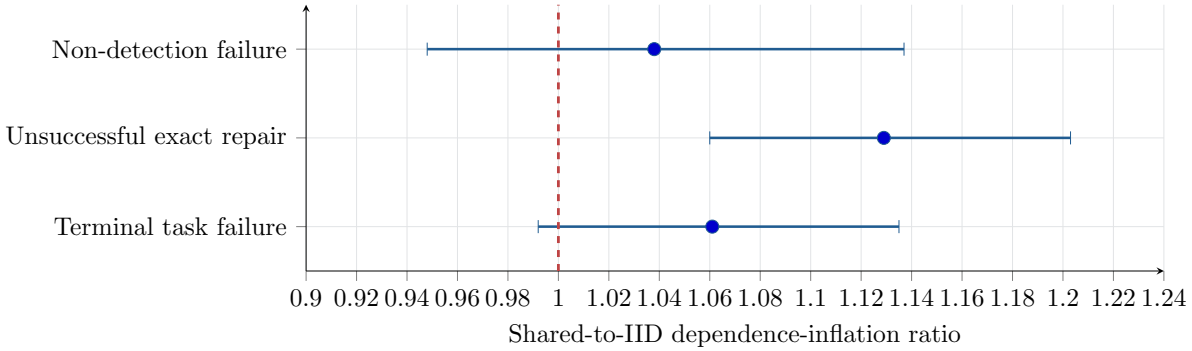

Figure 2: Primary dependence-inflation ratios and simultaneous intervals; the dashed line marks ratio 1. Only unsuccessful exact repair excludes one. Points and endpoints are the reported rounded values; no intermediate values are inferred.

These estimates come from the frozen implemented primary family, but they do not constitute an unqualified confirmatory success. The protocol required at least one simultaneous interval above one, use of its specified point-estimator weights, and passage of every predeclared gate. The implementation instead used equal item weights, and the required precollection calibration-support projection gate was omitted and retrospectively fails, as detailed below.

For comparison, the primary report gives terminal marginal failure rates of 48.58% in the clean condition, 50.82% under shared corruption, and 50.35% under IID corruption. These rates summarize average endpoint behavior rather than cross-endpoint dependence.

4.2 Weighting, support, and missingness

Table 2 presents three named postconfirmatory analyses of unsuccessful-repair failure dependence. The equal-mechanism and calibration-composition estimates retain the 1,677-item joint-exact primary support. Of the available sensitivities, equal-mechanism weighting is closest to the protocol's fixed equal mechanism, domain, and difficulty-band target; it yielded 1.128 [1.074, 1.182]. The condition-specific analysis instead uses 1,766 shared-arm items and 1,750 IID-arm items. Although all three intervals lie above one, they are unadjusted percentile intervals from specification analyses, not replications on an identical population or additions to the primary family.

Table 2: Named postconfirmatory unsuccessful-repair sensitivities. Intervals are unadjusted 95% percentile intervals and must not be substituted for the primary simultaneous interval.

Sensitivity	DI_{shared}	DI_{iid}	Ratio [interval]
Equal-mechanism weights	2.475	2.195	1.128 [1.074, 1.182]
Calibration-composition weights	2.840	2.579	1.101 [1.049, 1.153]
Condition-specific support	2.469	2.189	1.128 [1.076, 1.182]

Across six named deterministic missing-cell scenarios, the unsuccessful-exact-repair failure-dependence ratio ranged from 1.127 to 1.141. This scenario range does not bound every missingness process. The conservative envelope over unrestricted binary completion was wider, 0.989–1.311, and included one; it is not a confidence interval. The named specifications were stable, but the result remains unidentified under arbitrary outcome-dependent missingness (CMFC D1 Publication Closeout, 2026).

Non-detection remained unconfirmed under both alternate weighting schemes. A postconfirmatory terminal-failure interval lay slightly above one, but it cannot upgrade the null-containing primary terminal interval because it uses a different population, interval construction, and inferential family.

4.3 Average failure and the all-eight tail

For unsuccessful repair, mean route failure changed from 59.23% under IID assignment to 59.91% under shared assignment. The paired change was 0.68 percentage points with interval [−0.12, 1.45], which crossed zero. At the separate "at least half failed" threshold ($\geq 4/8$, not strict majority), failure was 69.65% under IID and 68.99% under shared assignment, a slight decrease. The all-eight summary moved in the other direction: unsuccessful repair occurred in 77 of 1,677 IID item families (4.59%) and 117 of 1,677 shared item families (6.98%). This corresponds to an absolute change of 2.39 percentage points [0.95, 3.82] and a ratio of 1.519 [1.184, 1.984] (CMFC D1 Publication Closeout, 2026; CMFC D1 Errata, 2026).

Figure 3 places these three descriptive percentage-point differences on a common axis without equating them. The mean and all-eight intervals are postconfirmatory, percentile, and unadjusted. The at-least-half point is a subtraction of the reported rates, so no interval is plotted for it. The divergence is specific to the all-eight summary; it does not establish a general upward shift in every upper-tail measure or a latent tail mechanism.

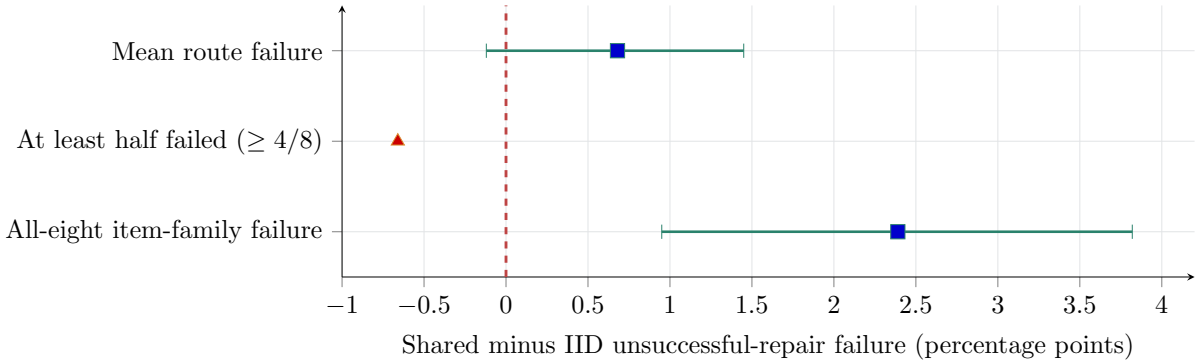

Figure 3: Postconfirmatory shared-minus-IID percentage-point contrasts for three unsuccessful-repair summaries. The dashed line marks no difference. Mean and all-eight points have reported unadjusted percentile intervals. The at-least-half point (triangle) is the descriptive difference derived from reported rates, with no interval inferred.

For terminal failure, the all-eight ratio was 1.229 [0.920, 1.658]. Non-detection produced zero all-eight failures in either corrupt condition, leaving its tail ratio non-estimable without a prohibited pseudocount. The extreme-tail pattern is therefore confined to unsuccessful exact repair and does not show a fleet-wide increase across outcomes.

4.4 Secondary closeout analyses

Under its frozen gates, the corrected observable-stage analysis suppressed DI and all-eight ratios for every preterminal raw stage. Terminal answer was the only reportable raw DI stage, at 1.064451 [1.012918, 1.119762], but this postconfirmatory interval cannot replace the null-containing primary terminal interval (CMFC D1 Stage V2, 2026). Suppression indicates that the earlier stages were not estimable under the gate; it does not imply that dependence began with the terminal answer. The analysis could not locate an observable onset, and emitted fields do not reveal a cognitive sequence.

The historical infrastructure analysis was a separate descriptive, observational, and matching-sensitive study. Equally optimal matchings changed joint-failure counts and, in one window, reversed the sign of an excess-co-unavailability retry contrast. Final pairwise dependence was classified as non-identified (CMFC Infrastructure V3, 2026). The analysis identifies neither provider independence nor a causal retry effect, and neither secondary analysis tested a repair intervention.

4.5 The omitted projection gate

The protocol required a precollection projection of at least 20 events and 20 non-events for every endpoint, corrupt condition, and primary outcome (CMFC D1 Protocol, 2026), but the collection gate did not implement the check. Reconstruction from the frozen calibration population fails it: the Cohere route projects zero unsuccessful-repair successes in both corrupt arms, and some non-detection cells fail as well. On realized joint support, Cohere recorded only three shared-arm and four IID-arm exact-repair successes (CMFC D1 Publication Closeout, 2026).

The deviation is substantive because sparse endpoint-outcome cells can make a dependence estimate sensitive even when the aggregate matrix is numerically defined. Stability in the named sensitivities cannot retroactively satisfy a required design gate. At the same time, failure of the projection does not render every estimate meaningless. The result is properly classified as an

unsuccessful-repair failure-dependence interval above one under the frozen implemented analysis from a study that did not meet its full confirmatory-success rule.

5 Discussion

Within this finite synthetic eight-route population, sharing one complete corruption variant increased estimated dependence inflation for unsuccessful exact-repair failure relative to IID-with-replacement assignment. The corresponding primary intervals for non-detection and terminal failure included one. Because the outcomes were scored separately, this contrast does not show that routes first flagged a problem, converged on the same wrong repair, or followed a common internal sequence.

The average, at-least-half, and all-eight summaries tell different parts of the result. Mean repair failure changed by less than one percentage point and its interval crossed zero, while the at-least-half rate moved slightly downward. All-eight failure, by contrast, increased by 40 item families, from 77 to 117. An average alone would miss that all-member pattern. The all-eight statistic remains postconfirmatory and does not establish deployment significance, but it adds information that neither the mean nor the intermediate threshold contains.

These findings also narrow what can be inferred from calls for "model diversity." Classical ensemble work explains why disagreement can help, while empirical comparisons show that diversity measures do not map cleanly to ensemble accuracy (Krogh and Vedelsby, 1994; Kuncheva and Whitaker, 2003). In language-model systems, different interfaces, providers, or model labels likewise do not guarantee independent errors (Kim et al., 2025; Kim, 2026). Here, complete shared corruption produced higher exact-repair failure dependence than IID-with-replacement draws from the same pool. The design cannot determine whether that contrast was mediated by false value, target position, severity, relevance, salience, packet identity, or latent item difficulty.

Nor did the experiment test a remedy. Separate evidence stores, diversified repair prompts, different providers, and communication restrictions remain hypotheses for future intervention studies rather than supported prescriptions. The stage decomposition cannot locate a cognitive onset because emitted fields are checkpoints in a response contract and the frozen sparsity gate suppresses every preterminal DI ratio. The historical infrastructure analysis is also observational and matching-sensitive, precluding provider attribution or a causal estimate of retry effects.

Evaluation programs should distinguish route-level failure, failure of a chosen aggregation rule, cross-route concentration on the same item, and the effects of controlled interventions on that concentration. This study examines the third question for a specific complete-variant sharing manipulation and reports the first descriptively; it does not answer the intervention question.

6 Limitations

Scope and measurement. The study population is defined by its synthetic tasks, corruption pools, route deployments, output contract, and collection epoch. It includes arithmetic, constraint, extraction, and tool-call items, but not code, long-horizon recovery, or real-world corruption prevalence, so the results should not be generalized to arbitrary agents, domains, or future route versions. Exact repair is an operational construct that requires the designated replacement object: a response may still be useful without satisfying that contract, and a terminally correct answer can bypass repair. A flag likewise records observable output rather than internal awareness. The complete corruption variant also bundles false value, target position, severity, relevance, salience,

packet identity, and latent item difficulty, leaving those potential mediators unresolved. IID assignment used replacement rather than guaranteed-distinct draws, and the bounded shareable artifacts omit the promised realized collision summary. With no repeated behavioral generations for each route-item-condition cell, persistent route tendencies cannot be separated from generation-level stochasticity.

Inference and protocol. The item-family bootstrap holds fixed the eight routes, construction and corruption pools, retained support, and observed stratum composition. It describes stability to resampling retained item families within strata, not uncertainty over routes, providers, future model versions, or real-world tasks. Empirical coverage of the selected simultaneous interval rule cannot be reverified without the design-simulation bytes. The frozen primary point estimator also weighted the 1,677 retained item families equally, whereas the protocol specified fixed equal weights over mechanisms, domains, and difficulty bands. Stratified bootstrap resampling preserved observed stratum counts but did not supply those point-estimator weights; the closest retained equal-stratum sensitivity was postconfirmatory and used an unadjusted percentile interval. In addition, the omitted calibration-support projection gate retrospectively fails. Sparse realized cells, including the Cohere repair-success counts, make the deviations substantive rather than clerical. The implemented estimates are numerically defined and the named sensitivities are stable, but the study is not a fully protocol-conforming confirmation.

Support, missingness, and infrastructure. The primary analysis uses 1,677 item families complete across all three conditions. The condition-specific analysis uses 1,766 shared and 1,750 IID items, while stage V2 uses 1,716 corrupt-paired and 1,758 clean items. Postconfirmatory percentile intervals do not inherit the simultaneous familywise status of the primary intervals, and the unrestricted binary-completion envelope is not a confidence interval and includes one. For the historical infrastructure question, delivery timing, scheduling, routing, and retries were not randomized, and pairwise final-state results change under equally optimal matchings. The analysis identifies neither provider independence nor causal retry effects.

Reproducibility. The bounded shareable evidence package, which has not been externally released, contains protocols, reports, aggregate JSON, selected source and tests, reviews, tables, manifests, and additive closure corrections. It excludes restricted prompts, responses, item and gold state, credentials, provider payloads, verdict rows, event databases, restricted item populations, and three local reproduction artifacts containing absolute builder-host paths. Hashes identify absent dependencies without supplying their bytes. For this draft, primary aggregate evidence was reopened and reported quantities were checked against retained outputs, but collection, bootstrap inference, calibration, and row-level scoring were not rerun.

7 Conclusion

In the frozen implemented equal-item analysis of this finite synthetic eight-route study, sharing the same complete corruption variant produced a simultaneous interval above one for unsuccessful exact-repair failure dependence. The corresponding primary intervals for non-detection and terminal failure included one. Named postconfirmatory sensitivities also yielded unsuccessful-repair failure-dependence estimates above one, although the unrestricted binary-completion envelope included the null. The postconfirmatory summaries diverged: all-eight failure increased, the interval for the mean included zero change, and the at-least-half threshold did not increase.

These results do not amount to an unqualified protocol-conforming confirmation. The point-estimator weighting departed from the protocol text, and the required calibration-support projection gate was omitted and retrospectively fails. The secondary analyses do not establish an internal mechanism, provider attribution, a causal retry effect, or a validated remedy.

Data, code, and evidence availability

The evidence base for this manuscript is the read-only bounded package `cmfc-d1-follow-up-v2`, archive SHA-256 `88bb3e0ba726ccde4771a5f13b5f67017097857d74cf31896e3d5713832773d9`. File references below are archive-relative. The empirical values reported here appear in `primary-d1-v1/analysis.json`, `artifacts/d1-followups/publication-closeout/summary.json`, `artifacts/d1-followups/stage-cascade-v2/summary.json`, and `artifacts/infrastructure-dependence-v3/summary.json`. The original bounded package is not redistributed in full and is neither a complete executable environment nor an unrestricted dataset. Selected report-level results, a methods note, and this paper are available at https://www.future-shock.ai/research/when-models-fail-to-repair. The public result downloads reproduce reported aggregates, not raw event records or a complete analysis environment. Restricted upstream artifacts remain withheld. Preparation of this paper involved no new scientific experiment, model collection, provider call, or empirical reanalysis.

Declarations

- **Authorship and contributions:** Nicholas Zinner and Beacon Bot, Future Shock. Nicholas Zinner directed the research and authorized this release. Beacon Bot assisted with evidence review, literature retrieval, drafting, revision, and document production.
- **Funding and conflicts of interest:** Self-funded by Nicholas Zinner; no external funding or competing interests to declare.
- **Evidence and data rights:** This release presents synthetic-task results and report-level evidence. Restricted prompts, responses, gold state, provider payloads, and event databases are not released. No claim of external ethics certification is made.
- **Licensing and redistribution:** Publication of this paper does not grant redistribution rights to restricted or referenced upstream artifacts.
- **AI assistance and responsibility:** Human authors remain responsible for scientific interpretation, declarations, and publication decisions. AI-assisted review is not independent empirical replication.
- **Publication state:** Public research working paper, not peer reviewed. The briefing, methods, and report-level result downloads are available at https://www.future-shock.ai/research/when-models-fail-to-repair.

A Retained route identifiers

The retained route IDs are reproduced verbatim from the primary support inventory:

1. anthropic-claude-haiku45-anthropic
2. cohere-command-r7b-cohere
3. deepseek-v4-flash-streamlake-fp8
4. google-gemini31-flash-lite-aistudio
5. ibm-granite41-8b-coreweave-bf16
6. mistral-small-2603-mistral
7. openai-gpt41-mini-openai
8. zai-glm53-flash-zai-fp8

B Evidence classes and archive handles

Table 3: Evidence classes are kept separate because they establish different propositions.

Class	Canonical archive-relative handle
Primary protocol	primary-d1-v1/protocol.md
Operative calibration amendment	primary-d1-v1/amendment-0021.md
Primary analysis	primary-d1-v1/analysis.json
Primary report	primary-d1-v1/report.md
Postconfirmatory sensitivity	artifacts/d1-followups/publication-closeout/summary.json
Sensitivity errata	docs/results/d1-publication-sensitivity-closeout-errata-2026-09-02.md
Corrected closure inventory	artifacts/d1-followups/publication-closeout/CLOSURE_V2_SHA256SUMS
Closure provenance correction	docs/results/d1-publication-closeout-closure-provenance-correction-2026-09-03.md
Corrected observable stages	artifacts/d1-followups/stage-cascade-v2/summary.json
Stage source-binding supplement	docs/results/d1-stage-cascade-v2-provenance-supplement-2026-09-02.md
Historical infrastructure	artifacts/infrastructure-dependence-v3/summary.json

The version-2 publication-closeout checksum inventory supersedes only the version-1 verification inventory. Its expected version-1 review-byte mismatch remains part of the record, and no analytical output changed. The prospective five-file executed-code map for the stage analysis omitted six transitive local modules. The later supplement records their closure-time hashes and an independent numerical reconstruction without making that prospective map retroactively exhaustive.

C Successor fixture-validation note

A separately packaged standard-library successor API reports 12 retained test methods and 19 frozen synthetic fixture results. The fixtures cover fleet metrics, log-DI edge cases, weighted DI, binary missingness bounds, projection classification, local response scoring, and canonical serialization. Its `verification.json` binds nine source/test files and three sibling outputs, while the completed closeout and package manifest bind all four run002 outputs (CMFC A2F Run002, 2026). During the prior bounded intake, all 19 fixture results were reconstructed independently from included oracles without importing or executing the supplied implementation. This evidence concerns the software interface and synthetic fixture arithmetic only. It did not reanalyze empirical records, rerun bootstrap inference or calibration, rescore historical outcomes, or resolve missing data.

References

Krogh, A. and Vedelsby, J. (1994). Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems 7*, 231–238. https://proceedings.neurips.cc/paper_files/paper/1994/hash/b8c37e33defde51cf91e1e03e51657da-Abstract.html

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, LNCS 1857. Springer. https://doi.org/10.1007/3-540-45014-9_1

Kuncheva, L. I. and Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51, 181–207. https://doi.org/10.1023/A:1022859003006

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations*. https://arxiv.org/abs/2203.11171

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv:2110.14168*. https://arxiv.org/abs/2110.14168

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. (2023). Let's verify step by step. *arXiv:2305.20050*. https://arxiv.org/abs/2305.20050

Jiang, D., Ren, X., and Lin, B. Y. (2023). LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 14165–14178. https://doi.org/10.18653/v1/2023.acl-long.792

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, 11733–11763. https://proceedings.mlr.press/v235/du24e.html

Estornell, A. and Liu, Y. (2024). Multi-LLM debate: Framework, principals, and interventions. *Advances in Neural Information Processing Systems 37*. https://proceedings.neurips.cc/paper_files/paper/2024/hash/32e07a110c6c6acf1afbf2bf82b614ad-Abstract-Conference.html

Peng, M., Chen, N., Tang, J., and Li, J. (2025). How does misinformation affect large language model behaviors and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 13711–13748. https://aclanthology.org/2025.acl-long.674/

Kim, E., Garg, A., Peng, K., and Garg, N. (2025). Correlated errors in large language models. *International Conference on Machine Learning*; arXiv:2506.07962. https://arxiv.org/abs/2506.07962

Kim, D. (2026). Are diversity metrics measuring diversity? A capability-controlled audit of majority-vote gain in LLM ensembles. *arXiv:2607.20768*. https://arxiv.org/abs/2607.20768

Little, R. J. A. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, 3rd ed. Wiley. https://doi.org/10.1002/9781119482260

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review*, 80(2), 319–323. https://ideas.repec.org/a/aea/aecrev/v80y1990i2p319-23.html

CMFC D1 (2026). D1 clean-corruption detection and recovery protocol v1. `cmfc-d1-follow-up-v2/primary-d1-v1/protocol.md`.

CMFC D1 (2026). Amendment 0021: pooled calibration gate and code exclusion. `cmfc-d1-follow-up-v2/primary-d1-v1/amendment-0021.md`.

CMFC D1 (2026). D1 clean-corruption detection and recovery study. `cmfc-d1-follow-up-v2/primary-d1-v1/report.md`; load-bearing values in `primary-d1-v1/analysis.json`.

CMFC D1 (2026). D1 publication-sensitivity closeout. `cmfc-d1-follow-up-v2/docs/results/d1-publication-sensitivity-closeout-2026-09-02.md`; retained output `artifacts/d1-followups/publication-closeout/summary.json`.

CMFC D1 (2026). D1 publication-sensitivity closeout errata. `cmfc-d1-follow-up-v2/docs/results/d1-publication-sensitivity-closeout-errata-2026-09-02.md`.

CMFC D1 (2026). D1 observable-stage cascade v2. `cmfc-d1-follow-up-v2/docs/results/d1-stage-cascade-2026-09-02-v2.md`; retained output `artifacts/d1-followups/stage-cascade-v2/summary.json`.

CMFC D1 (2026). Infrastructure dependence v3 corrective analysis. `cmfc-d1-follow-up-v2/docs/results/infrastructure-dependence-2026-09-03-v3.md`; retained output `artifacts/infrastructure-dependence-v3/summary.json`.

CMFC A2F (2026). Run002 narrow closeout and verification. `a2f-m1/evidence/HANDOFF-final.md`; `a2f-m1/outputs/m1-run-002/verification.json`.