When Do Diverse Models Fail Together?

Difficulty, Shared Evidence, and Common Shocks
Across Language-Model Fleets

Nicholas Zinner Beacon Bot

Future Shock (future-shock.ai)

August 2026

Abstract

Across 27,842 exact common retrospective items, median raw chance-adjusted agreement (CAPA) was 0.498, but median residual correlation after a two-parameter item-response adjustment was −0.018. Most of the broad co-failure signal was therefore consistent with shared item difficulty rather than a large average residual model common cause. A prospective ten-endpoint panel retained small positive cross-family residual dependence ($\beta = 0.068$, 95% cluster CI $[0.039, 0.093]$).

Empirical failure clusters showed almost no alignment with named developer families: adjusted Rand index 0.012 retrospectively and −0.071 prospectively. The directional prediction that evidence-path diversification would outperform a fixed endpoint substitution was rejected in both confirmations ($\Delta = -0.460$ $[-0.589, -0.327]$ and -0.150 $[-0.182, -0.119]$); the endpoint substitutions changed a bundle of model, family, provider, size, and precision.

Recoverable common shocks produced one modest legacy evidence signal at degraded support; both contemporary intervals included unity, and neither instruction-anchor contrast excluded unity. Separately, a severe semantic instruction-conflict stress amplified median joint failure 4.337-fold $[3.358, 5.885]$. That stress demonstrates a distinct synchronization mechanism, not its general prevalence in deployment. The results replace a universal model-correlation story with a conditional account of difficulty, endpoint composition, input paths, and shock mechanism.

1 Introduction

A fleet of language models is not automatically a fleet of independent risks. Models may share training material, evaluation difficulty, retrieval systems, prompts, providers, or upstream evidence. Conversely, raw agreement on hard items can be large even when residual failures are nearly independent. The deployment question is therefore conditional: how much independence remains after difficulty is controlled, and how much is lost when models receive a common erroneous input?

We study that question with binary, deterministically scored outcomes and keep four quantities separate throughout: marginal fleet loss, absolute fleet volatility, chance-adjusted dependence, and benchmark-relative effective diversity. The design progresses from broad retrospective measurement to fresh prospective items, evidence-path and endpoint contrasts, adversarial transfer, orchestration topology, and finally a difficulty-controlled shared-versus-independent shock experiment.

The paper makes three contributions. First, we decompose raw co-failure into an item-difficulty component and a residual-dependence component across two distinct endpoint populations, showing

that the former dominates in a large retrospective panel. Second, we introduce a paired within-item common-shock design that isolates whether sharing the *same* erroneous input increases fleet failure dependence beyond severity-matched but independently assigned errors. Third, we measure adversarial instruction conflict as a qualitatively separate synchronization mechanism, providing a bridge between ordinary co-failure and strong semantic attack transfer.

2 Related Work

Correlated model errors. Kim et al. [Kim et al., 2025] conducted a large-scale empirical evaluation of error correlation across over 350 LLMs on two leaderboard datasets and a resume-screening task, finding that models agree on the wrong answer far more often than chance would predict and that larger, more capable models exhibit higher correlation. They identify unequal question difficulty as a limitation of wrong-answer agreement but do not perform an item-response residual decomposition. Our retrospective analysis likewise finds substantial raw coincidence (CAPA ≈ 0.50), then models item difficulty and discrimination and finds median *residual* correlation near zero.

Model similarity and oversight. Goel et al. [Goel et al., 2025] introduced CAPA as a chance-adjusted measure of probabilistic agreement and documented that model mistakes become more similar with increasing capabilities, identifying risks for AI oversight. Their Appendix D.2.2 bins questions by the fraction of models answering correctly, reports CAPA as mostly stable across hardness bins, and concludes that question difficulty is not a significant confounder for the capability–similarity trend. Our result does not ignore that check: it uses item-level 2PL difficulty and discrimination terms, then measures residual pair dependence. The near-zero retrospective median shows that coarse hardness stratification and item-response residualization need not yield the same interpretation. Differences in estimand, population, pair support, and adjustment resolution are plausible moderators and should be tested in a direct methodological replication.

Consensus as correlated evidence. Denisov-Blanch et al. [Denisov-Blanch et al., 2026] demonstrated that crowd-wisdom strategies fail for LLM truthfulness because shared training data and inductive biases produce correlated outputs that increase consensus without improving correctness. Our common-shock experiment directly manipulates one channel of this shared-input problem—recoverable evidence errors and instruction anchors—while holding task difficulty fixed.

N-version independence. Knight and Leveson [Knight and Leveson, 1986] experimentally rejected the independence assumption of N-version programming in conventional software, finding statistically significant coincident failures across independently developed program versions. Our study extends this line of inquiry to language-model fleets, where "versions" differ not only in implementation but in training data, architecture, and provider infrastructure.

Algorithmic monoculture. Kleinberg and Raghavan [Kleinberg and Raghavan, 2021] formalized algorithmic monoculture as a systemic risk when many decision-makers deploy similar models. Bommasani et al. [Bommasani et al., 2022] identified homogenization risks in foundation models. Our evidence-path and endpoint-substitution results show that endpoint diversity and input diversity are distinct system properties, consistent with these theoretical concerns.

Item response theory. We use two-parameter logistic (2PL) IRT models [Lord and Novick, 1968, Embretson and Reise, 2000] to decompose item-level difficulty from model-level ability in the retrospective panel. The 2PL residual correlation serves as our primary measure of dependence after difficulty adjustment.

Bootstrap inference. Our simultaneous inference uses the studentized bootstrap with cluster resampling [Cameron et al., 2008, Davison and Hinkley, 1997], extended to a max-absolute-statistic procedure for joint coverage across two primary contrasts.

3 Study Design

3.1 Populations and Hierarchy

The study comprises four endpoint populations, analyzed separately and never pooled:

1. **Retrospective panel:** 395 models on 27,842 exact common items from public leaderboard and evaluation data.
2. **Prospective direct panel:** 10 fixed routed endpoints on 1,500 freshly constructed, machine-scored items (1,495/1,500 retained at $\geq$99.67% exact common support).
3. **Legacy common-shock confirmation:** 9 endpoints (originally 10; Qwen3-30B/Alibaba excluded before outcome access at a frozen infrastructure gate) on 800 fresh items $\times$ 6 conditions.
4. **Contemporary common-shock panel:** 6 endpoints from 6 developer families on 100 items $\times$ 6 conditions.

The retrospective execution reported its selection audit separately. It retained 3,253 support-qualified model pairs among 77,815 possible unordered pairs (4.18%). The Stage 1 public bundle does not include the event-level pair-support matrix needed to regenerate that selection count independently.

The legacy common-shock execution underwent a prospectively amended transport-recovery protocol after two prior infrastructure-gate failures. The recovery retained all terminal cells and reopened only formally retryable transport errors with a five-attempt ceiling under reduced concurrency. This amendment was selected outcome-blind but constitutes a protocol change.

3.2 Legacy Common-Shock Endpoints

The nine legacy endpoints span seven developer families and seven inference providers:

Table 1: Legacy nine-endpoint common-shock fleet.

Endpoint	Developer Family	Provider
GPT-4.1 Nano	OpenAI	OpenAI
GPT-4.1 Mini	OpenAI	OpenAI
Gemini 2.5 Flash-Lite	Google	Vertex EU
Claude 4.5 Haiku	Anthropic	Anthropic
Llama 3.1 8B	Meta	Groq
Llama 3.3 70B	Meta	Groq
Mistral Small 3.2 24B	Mistral	Mistral EU
Qwen3 8B	Alibaba	Alibaba
DeepSeek V3-0324	DeepSeek	Crusoe

3.3 Contemporary Endpoints

The six contemporary endpoints span six developer families: DeepSeek V4 Flash (DeepInfra), MiniMax M3 (MiniMax), Kimi K2.7 (Inceptron), Nemotron 3.5 (CoreWeave), Qwen3.8 27B (Chutes), and GLM-5.3 (ZAI). These are the frozen route labels recorded by the package, not independently validated claims about the underlying model weights.

Degeneracy disclosure. Claude 4.5 Haiku failed every item (100% failure) under both anchor conditions (A_{shared} and $A_{\text{independent}}$) in the legacy fleet, rendering it a zero-variance endpoint for the instruction-anchor contrast. The anchor DI contrast effectively operates on 8 nondegenerate endpoints. Nemotron 3.5 failed 97–100% of items across all conditions in the contemporary fleet (100% under both evidence conditions), making it degenerate for the evidence contrast and near-degenerate elsewhere. These degeneracies reduce effective fleet size and are not masked by equal-weight averaging.

3.4 Common-Shock Conditions and Estimands

Each item is assigned six condition variants:

Table 2: Six experimental conditions for the common-shock study.

Condition	Definition
O (Original)	Original task without reviewer proposal or evidence packet.
N (Neutral)	Length- and format-matched answer-preserving rewrite.
A_{shared}	Every endpoint receives the same wrong reviewer proposal.
A_{indep}	Endpoints receive different wrong proposals from the same pool.
E_{shared}	Every endpoint receives the same one-error evidence variant.
E_{indep}	Endpoints receive different one-error variants from the same pool.

Models are told that reviewer notes and evidence are untrusted and may contain one error; the operative task is always to derive and return the correct answer.

Primary estimands. Let $Y_{imc} \in \{0, 1\}$ denote binary failure for item i, endpoint m, condition c. With equal weights $w_m = 1/M$:

- *Fleet variance:* $V_c = \mathbf{w}^\top \Sigma_c \mathbf{w}$
- *Diagonal variance:* $D_c = \sum_m w_m^2 \operatorname{Var}(Y_{imc})$
- *Dependence inflation:* $\mathrm{DI}_c = V_c / D_c$
- *Primary contrasts:* $\theta_A = \log(\mathrm{DI}_{A_{\text{shared}}}/\mathrm{DI}_{A_{\text{indep}}}), \quad \theta_E = \log(\mathrm{DI}_{E_{\text{shared}}}/\mathrm{DI}_{E_{\text{indep}}})$

Positive values indicate that sharing the same error increased dependence relative to distributing severity-matched errors independently. We report $\exp(\theta)$ as the DI ratio. Mean loss, absolute volatility ratio (AVR), and tail-event rates are separate estimands and are not collapsed into DI.

3.5 Inference

Primary common-shock intervals use a 9,999-resample studentized max-absolute-statistic bootstrap, clustering by item family within frozen domain and predicted-difficulty strata. The max-absolute-statistic procedure provides simultaneous coverage for both θ_A and θ_E.

Singleton-stratum correction. In the contemporary panel, two singleton-cluster strata are treated as self-representing certainty strata under a documented postcollection implementation correction. This correction was made after legacy results existed and is therefore not described as outcome-blind. It changes bootstrap execution but no contemporary point estimate, item, verdict, stratum, or estimand.

3.6 Missingness and Support Quality

Infrastructure failures (transport errors, HTTP 429/503, ambiguous delivery, route mismatch, scorer failure) are never recoded as model failures. Primary contrasts use exact matched support—only items with accepted responses from all endpoints under both conditions of a contrast.

Table 3: Support quality tiers for common-shock contrasts.

Tier	Requirements
Confirmatory quality	$\geq$90% matched support overall and $\geq$80% per domain
Estimable degraded	$\geq$500/800 overall and $\geq$60% per domain; bounds required
Non-estimable	Below degraded floor

Sharp binary-fill bounds cover mean loss and fleet tail events under all possible resolutions of missing cells. These bounds make no distributional assumption about missingness.

4 Base Results: Difficulty and Residual Dependence

4.1 Retrospective Panel

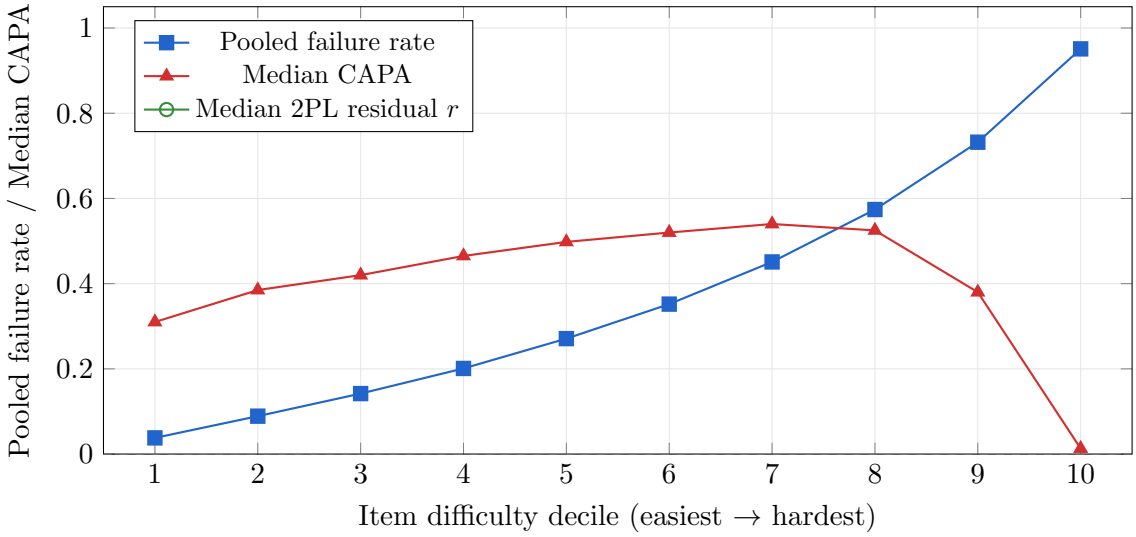

Figure 1: Difficulty decomposition in the 395-model retrospective panel. Raw CAPA tracks difficulty (blue) and peaks in the middle deciles, while 2PL residual correlation (green) is near zero across all deciles. The hardest decile (95.1% pooled failure) is ceiling-saturated; its low CAPA reflects near-universal failure, not benign tail independence.

Across 27,842 common items, median CAPA was 0.498 $[0.494, 0.503]$, while median 2PL residual correlation was -0.018. Figure 1 shows that CAPA tracks item difficulty and collapses at both extremes (where either almost everyone passes or almost everyone fails), while the residual correlation remains near zero throughout. The hardest decile reached 95.1% pooled failure—a near-universal-failure ceiling that does not identify tail independence or support a parametric tail copula.

Vendor taxonomy aligned weakly with empirical clustering: the adjusted Rand index was 0.012 across 127 models in 6 eligible multi-model families.

4.2 Prospective Direct Panel

The ten-endpoint prospective panel retained 1,495/1,500 items at exact common support. Cross-family residual dependence was small but positive:

Table 4: Diversification axis betas from the prospective panel. Each axis is a separate frozen execution and they are not pooled.

Axis	Manipulation	β	95% Cluster CI	Support
A0	Stochastic draws within endpoint	0.912	$[0.891, 0.942]$	273/300
A1	Private scaffolds within endpoint	0.840	$[0.764, 0.892]$	300/300
A2	Size/class within family	0.001	$[-0.021, 0.055]$	1495/1500
A3	Different model families	0.068	$[0.039, 0.093]$	1495/1500
A4	Evidence paths (country, primary)	0.539	$[0.407, 0.667]$	448/450
A4	Evidence paths (city, secondary)	0.640	$[0.613, 0.668]$	3263/3600

The cross-family axis (A3) shows residual dependence that is statistically distinguishable from zero but small in magnitude, consistent with modest shared difficulty structure that the 2PL model does not fully capture, or genuine residual co-failure.

4.3 Evidence Paths vs. Endpoint Substitution

The directional prediction that evidence-path diversification would exceed the benefit of a fixed endpoint substitution was rejected in both populations. The country contrast was $\Delta_{H4} = -0.460$ $[-0.589, -0.327]$; the city contrast was -0.150 $[-0.182, -0.119]$. The endpoint substitution jointly changes model identity, family, size, provider, and precision and is therefore not a causal family-only effect.

5 Adversarial Transfer and Orchestration

These experiments measure mechanisms distinct from both ordinary co-failure and the controlled common-shock study and are not pooled with either.

Semantic instruction conflict. A harmless semantic instruction-conflict attack amplified median pairwise joint failure 4.337-fold $[3.358, 5.885]$ (128/150 matched complete items). Token suffixes produced 0.826 $[0.642, 1.086]$. The semantic attack exploits a qualitatively different synchronization channel—models collectively follow a plausible but adversarially crafted reframe of the task—that is far more potent than the recoverable one-error shocks in the common-shock experiment.

Orchestration topology. In a matched architecture battery, role-position effective diversity was 4.533 $[3.096, 6.668]$ for parallel proposals and 1.580 $[1.347, 2.036]$ for chains. The paired chain-minus-parallel final-failure difference was -0.060 $[-0.180, 0.058]$.

Mechanism separation. The adversarial instruction-conflict result (4.337×) and the controlled common-shock evidence result (1.095×) look substantially different, but they do not share an estimand. The adversarial attack replaces the operative task instruction with a semantically coherent wrong directive, while the common-shock manipulations preserve the correct task instruction and inject a single recoverable error in supporting material. They address different deployment threat models and should not be collapsed into a single "correlation" number.

6 Common-Shock Experiment

6.1 Design Summary

The common-shock experiment crosses 800 underlying items from five balanced domains (structured fact extraction, quantitative table reasoning, rule application, code/API-document reasoning, tool/procedure sequencing) with six conditions and nine legacy endpoints (48,000 analytical cells plus canaries). A separate 100-item contemporary panel crosses the same conditions with six endpoints (3,600 cells). Items, conditions, and assignments were frozen before any model call.

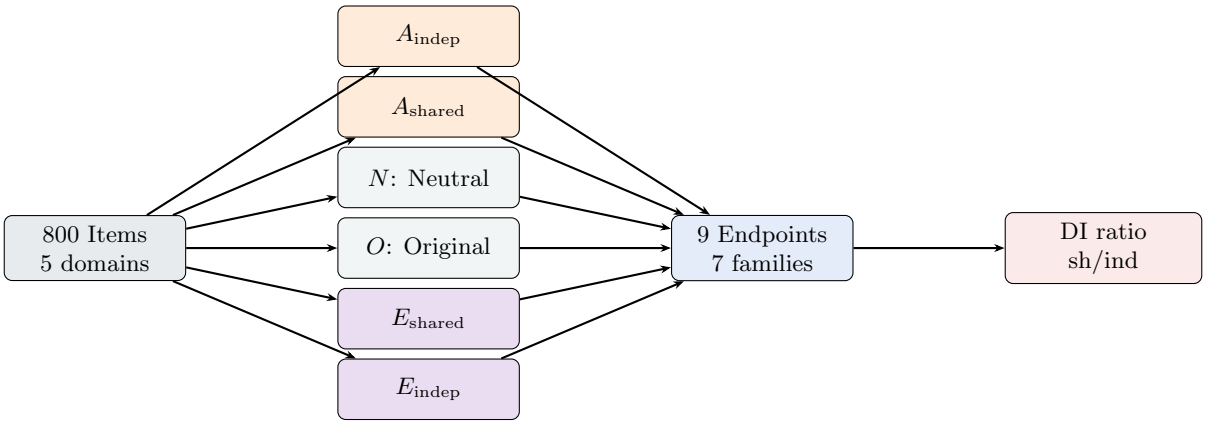

Figure 2: Common-shock experiment design. Each item is crossed with six conditions and dispatched to the full endpoint fleet. The primary estimand compares dependence inflation under shared versus independently assigned errors.

6.2 Primary Results

Table 5: Primary common-shock contrasts. Simultaneous 95% CIs from 9,999-resample studentized max-absolute-statistic cluster bootstrap. Support quality is displayed as the tier classification alongside matched support counts.

Population	Shock type	DI ratio	95% sim. CI	Support	Quality	$\Delta\bar{\mu}$
Legacy	Evidence	1.095	$[1.016, 1.181]$	665/800	*estim. degr.*	$+0.002$
Legacy	Anchor	0.979	$[0.925, 1.036]$	682/800	*estim. degr.*	$+0.003$
Contemp.	Evidence	1.007	$[0.888, 1.142]$	97/100	confirm. qual.	-0.007
Contemp.	Anchor	1.033	$[0.925, 1.153]$	94/100	confirm. qual.	-0.016

Legacy evidence-error contrast. The DI ratio of 1.095 $[1.016, 1.181]$ excludes unity, indicating that sharing the same recoverable evidence error modestly amplified fleet failure dependence relative to independently assigned errors of matched severity. However, this result comes from **estimable_degraded** support (665/800, with arithmetic at 71.9% and extraction at 76.9%—both below the 80% per-domain confirmatory floor). The CI lower bound is only 0.016 above 1.0 on the ratio scale. Mean-loss missingness bounds straddle zero ($[-0.014, +0.0115]$), so the mean-loss difference could change sign under different resolutions of missing cells.

Legacy instruction-anchor contrast. The DI ratio of 0.979 $[0.925, 1.036]$ includes unity. Sharing

the same wrong instruction anchor did not detectably increase fleet failure dependence relative to independently assigned anchors.

Contemporary contrasts. Both contemporary intervals span unity with wide margins, consistent with no effect but also consistent with effects as large as the observed legacy evidence signal. The contemporary panel has lower power (6 endpoints, 100 items) and is not a successful replication.

6.3 Condition-Level Risk Decomposition

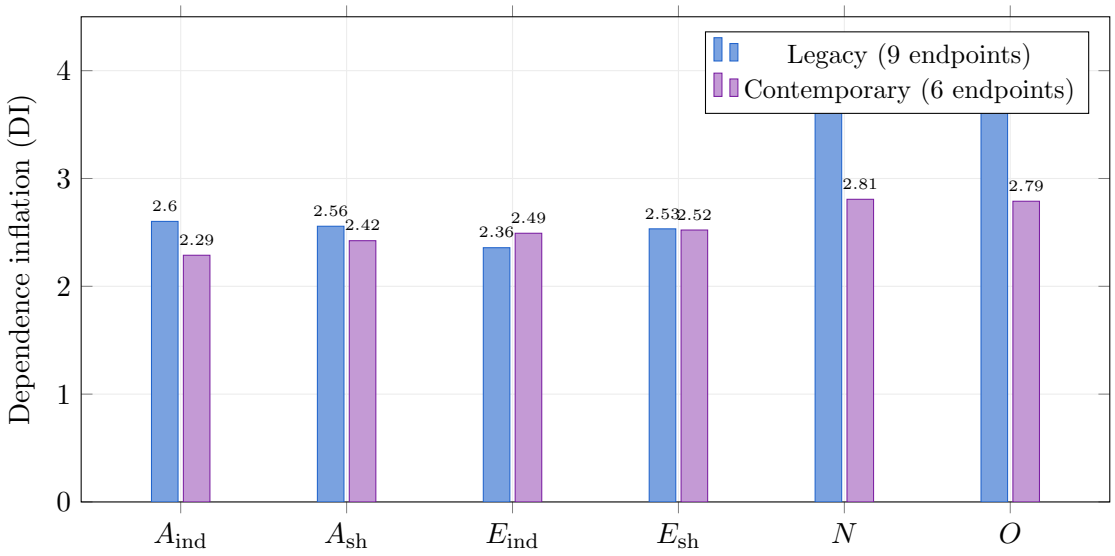

Figure 3: Condition-level dependence inflation across both populations. Original (O) and neutral (N) conditions show higher DI than treated conditions because treatment-induced errors increase marginal variances (the DI denominator) more than off-diagonal covariance. The primary contrasts compare shared versus independent within each treatment type, not treatment versus baseline.

6.4 Neutral Manipulation Check

The neutral rewrite was designed as a difficulty-preserving control. The original-to-neutral mean failure difference was -0.040 in the legacy population and -0.033 in the contemporary population. Both exceed the prespecified ± 0.03 equivalence margin, indicating that the neutral rewrite is systematically harder than the original.

The legacy dependence ratio $\text{DI}_N/\text{DI}_O = 0.995$ is within the equivalence band $[0.90, 1.10]$, suggesting that the transformation shifted marginal difficulty without altering the dependence structure. Neutral-check failure limits interpretation of total anchor-versus-neutral effects. It does not directly test the shared-versus-independent ratio, but it weakens confidence that the transformation pipeline changed only the intended feature. The primary contrast still compares conditions sharing the same underlying items and error type while varying error-identity assignment.

6.5 Contemporary Canary Absence

The contemporary panel has zero canary observations across all six endpoints and all three planned epochs (start, midpoint, end). The canary mechanism that was designed to detect mid-execution drift was entirely absent. No drift-detection diagnostic is therefore available for the contemporary execution. This limits assurance that endpoint behavior remained stable during collection.

6.6 Missingness and Sharp Bounds

Table 6: Condition-level missingness and sharp bounds (legacy fleet). Missing cells counts exclude the Qwen3-30B endpoint removed at the infrastructure gate.

Condition	Items	Missing	$\bar{\mu}$ range	Maj.-fail range	All-fail range
A_{indep}	723	92	$[0.529, 0.542]$	$[0.488, 0.506]$	$[0.016, 0.025]$
A_{shared}	733	86	$[0.532, 0.544]$	$[0.499, 0.514]$	$[0.014, 0.015]$
E_{indep}	723	89	$[0.539, 0.551]$	$[0.494, 0.515]$	$[0.016, 0.026]$
E_{shared}	717	95	$[0.537, 0.550]$	$[0.491, 0.509]$	$[0.016, 0.028]$
N (Neutral)	733	77	$[0.486, 0.497]$	$[0.453, 0.461]$	$[0.055, 0.064]$
O (Original)	736	83	$[0.447, 0.458]$	$[0.414, 0.423]$	$[0.034, 0.043]$

The legacy confirmation experienced 42,813 accepted cells, 334 ambiguous events, and 188 infrastructure errors out of 43,335 attempted cells. Terminal error types included ConnectError (1,825), HTTP 429 (312), ReadError (239), RemoteProtocolError (186), and TimeoutError (31). Missing cells are concentrated in transport-sensitive provider routes and are not distributed uniformly across endpoints or domains.

7 Diagnostics and Amendment History

7.1 Canary Drift

In the legacy fleet, all nine endpoints showed zero canary drift across start, midpoint, and end epochs (maximum range 0.0 across all endpoints except Qwen3 8B at 0.2, a one-item shift among five canaries per epoch). In the contemporary fleet, all canary observations are absent (zero observed per epoch per endpoint), as disclosed in Section 6.5.

7.2 Deployment Mix Illustrations

Table 7: Prospective deployment mix illustrations (base protocol). n_{eff} is a benchmark-relative covariance-diversification index. These are benchmark-relative illustrations, not capital requirements.

Mix	Endpoints	n_{eff}	Realized %
OpenAI within-family pair	2	1.98	99.2%
Llama within-family pair	2	1.20	60.1%
Qwen within-family pair	2	1.54	77.2%
Three-family fleet	3	2.06	68.6%
Seven-family fleet	7	2.34	33.5%

7.3 Amendment Chronology

The common-shock study underwent twenty named protocol amendments between contract freeze (2026-08-23) and analysis completion. Most are claimed outcome-blind and documented with frozen artifacts and SHA-256 hashes; the contemporary singleton-stratum correction explicitly is

not described as outcome-blind. The amendment chain is entirely self-attested through frozen artifacts; it is not independently witnessed or timestamped by a third party.

Amendment themes include exchangeable-null simulator and Monte Carlo calibration fixes; pilot restart, ambiguity, support-floor, and active-time changes; domain-tier, output-contract, route, arithmetic-mixture, and error-envelope remediations; confirmation activation, throughput, streaming, and transport-recovery changes; and contemporary streaming and singleton-stratum corrections.

The volume of amendments creates garden-of-forking-paths exposure. A reader cannot fully audit this chain from the paper alone; the full amendment documents are retained with the local evidence package.

8 Discussion

The combined evidence rejects two simple stories.

First, high raw co-failure is not sufficient evidence that diverse models share a large residual failure mode. In the 395-model retrospective panel, ordinary item difficulty explains most of the observed CAPA (0.498), leaving a median residual correlation of -0.018. This decomposition is not available from raw agreement statistics alone and requires a difficulty model such as 2PL IRT.

Second, low average residual dependence does not guarantee safety under shared inputs. The adversarial instruction-conflict result ($4.337\times$ amplification) shows that a particular common semantic attack can synchronize failures dramatically, while the controlled common-shock experiment estimates the effect of recoverable, severity-matched shared errors under moderate baseline difficulty.

Three distinct mechanisms. The study distinguishes three sources of apparent or real co-failure: (i) shared item difficulty, which dominates the retrospective CAPA; (ii) shared recoverable errors, which produced one modest, fragile, degraded-quality signal in the legacy evidence contrast; and (iii) adversarial instruction conflict, which operates through a qualitatively different channel and produced a substantially larger numerical amplification under a different estimand. Treating these as a single "correlation" number would conflate estimands with very different deployment implications.

Measurement implications. Reliability evaluation should report input-sharing topology alongside endpoint composition. A fleet of diverse models behind one prompt, one retriever, or one evidence cache retains endpoint diversity while potentially losing input diversity. The evidence-path results (Section 4.3) and the common-shock experiment both indicate that these two dimensions of diversity are not interchangeable.

Implications for insurability. The retrospective decomposition weakens the presumption that routine agent losses are undiversifiable because model-error correlation is universally enormous. It does not establish that routine agent fleets are insurable. The study contains no claims frequency, monetary severity, portfolio weights, policy terms, or long-horizon loss data. It replaces one blanket objection with a conditional question: which tasks, dependencies, and shared-state mechanisms concentrate loss in a particular exposure topology?

Proposed post-hoc transition analysis. The existing common-shock execution could support an exploratory error-induced transition analysis if the event-level estate is recovered. For endpoint

m, item i, and evidence condition c, define $S_{mic} = 1$ when the endpoint passed the neutral rewrite (N) and failed the evidence condition (E_c), then estimate marginal transition rates and cross-endpoint coincidence of S. This may measure correlated susceptibility after removing many baseline failures. It is not a clean error-detection rate: N is not a matched error-free evidence packet, and an endpoint can detect an error yet fail, or pass by ignoring the packet. The required endpoint-by-item cells are referenced by the larger 993-file evidence estate but are absent from the delivered 20-file package and Stage 1 bundle. No transition estimate is reported here.

We emphasize measurement implications rather than deployment recommendations. The DI ratio does not determine deployment loss without traffic weights, severities, and a decision policy. Equal endpoint weights are a scientific reference portfolio, not a production routing policy.

9 Limitations

1. **Finite populations.** All conclusions apply to the named endpoint panels, exact provider routes, deterministic task contracts, and collection epochs. Router aliases do not cryptographically prove weight identity.

2. **Degraded legacy support.** Both legacy common-shock contrasts are classified **estimable_degraded**, not confirmatory quality. The sole null-excluding result (evidence DI ratio 1.095, lower CI bound at 1.016) leaves little margin around unity. Direct sharp bounds for the dependence estimand are not included in the delivered package.

3. **Inconclusive contemporary panel.** The contemporary contrasts have wide intervals that include unity; they are not successful replications. The panel has lower power and does not confirm or disconfirm the legacy evidence signal.

4. **Failed neutral equivalence.** The neutral manipulation check failed in both populations (O–N mean difference -0.040 legacy, -0.033 contemporary; margin ± 0.03). The transformation pipeline itself introduces difficulty shifts.

5. **Zero contemporary canaries.** The drift-detection mechanism was absent for the contemporary execution. No temporal stability assurance is available.

6. **Claude anchor degeneracy.** Claude 4.5 Haiku is 100% degenerate under both anchor conditions, reducing the effective legacy anchor fleet to 8 endpoints.

7. **Nemotron evidence degeneracy.** Nemotron 3.5 is degenerate (100% failure) under both contemporary evidence conditions and near-degenerate (96–99%) elsewhere.

8. **Non-outcome-blind singleton correction.** The contemporary singleton-stratum correction was made after legacy results existed.

9. **Self-attested amendment chain.** The twenty protocol amendments are documented through frozen artifacts and hashes but not independently witnessed.

10. **Binary summaries.** CAPA, DI, and related statistics are binary-outcome summaries. They do not capture error severity, semantic category, or causal mechanism.

11. **Recoverable shocks only.** The common-shock manipulations are harmless, single-error, recoverable interventions. They do not span dose response, live retrieval, malicious jailbreaks, or long-horizon agent behavior.

12. **Incomplete historical provenance.** The exact collection-time code patch is unavailable for some early base executions, though all runs are hash-described.

10 Reproducibility Boundary

The delivered evidence package establishes internal consistency but not full independent reproducibility. Frozen plans, event ledgers, accepted payloads, deterministic verdicts, matrices, bootstrap outputs, source manifests, checksums, and completion receipts are retained locally. The package omits restricted raw Phase-1 bytes pending redistribution permission.

The first public bundle supports report-level inspection and receipt reconstruction. It does not reproduce the historical model calls or regenerate every analysis from raw event-level evidence. Re-executing the identical provider calls would require the same model weights, serving infrastructure, quantization, and routing state at collection time—conditions that cannot be guaranteed retrospectively.

The public release is limited to the HTML briefing, this PDF, a methodology-and-receipts route, and a bounded Stage 1 evidence bundle. It does not constitute an independently reproducible execution.

11 Disclosure and AI Contribution

Study posture. This is a finite-panel reliability study, not a model ranking, deployment rule, or demonstration that shared errors generally erase model diversity.

AI authorship. Beacon Bot (an AI research assistant built on language models) contributed to study design, item generation, code implementation, data collection, statistical analysis, and manuscript preparation under human direction and review. Nicholas Zinner directed the research program, specified the study contract, reviewed all protocol amendments, and is responsible for the scientific content and claims. No vendor endorsement is implied by the use of named model endpoints.

Trademark notice. All model, provider, and platform names mentioned in this paper are trademarks of their respective owners and are used solely for identification in a research context.

Funding and conflicts. No external funding was received for this study. The authors declare no competing interests.

References

Bommasani, R., Hudson, D. A., Adeli, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2022.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3):414–427, 2008.

Davison, A. C. and Hinkley, D. V. *Bootstrap Methods and their Application*. Cambridge University Press, 1997.

Denisov-Blanch, Y., Kazdan, J., Chudnovsky, J., Schaeffer, R., Guan, S., et al. Consensus is not verification: Why crowd wisdom strategies fail for LLM truthfulness. *arXiv preprint arXiv:2603.06612*, 2026.

Embretson, S. E. and Reise, S. P. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, 2000.

Goel, S., Strüber, J., Auzina, I. A., Chandra, K. K., Kumaraguru, P., Kiela, D., Prabhu, A., Bethge, M., and Geiping, J. Great models think alike and this undermines AI oversight. *ICML 2025 poster*; arXiv:2502.04313, 2025.

Kim, E., Garg, A., Peng, K., and Garg, N. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, 2025.

Kleinberg, J. and Raghavan, M. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

Knight, J. C. and Leveson, N. G. An experimental evaluation of the assumption of independence in multiversion programming. *IEEE Transactions on Software Engineering*, SE-12(1):96–109, 1986.

Lord, F. M. and Novick, M. R. *Statistical Theories of Mental Test Scores*. Addison-Wesley, 1968.

Wang, Y., Ma, X., Zhang, G., et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.