

A Second Model Is Not a Second Opinion: Peer Verification, Correlated Failure, and the Independence Variable Nobody Reports

Pranay M. Mahendrakar

SONYTECH

ORCID: 0009-0003-7224-029X

Life of Research

2026-08-21

Abstract

When a language model cannot reliably detect its own errors, the standard remedy is to have a second model check the first. That substitution is now the default across agent benchmarks, reward pipelines and safety evaluations, and it rests on an assumption that is rarely stated and that, in the agent-verification papers surveyed here, is never measured: that the checker's errors are independent of the checked system's. This paper separates two things the phrase "peer verification" conflates - adding an evaluator instance, and adding an independent signal - and argues that only the second is what the justification requires. Read through that distinction, a literature that appears to disagree about whether judge panels help turns out to be consistent: aggregation reliably reduces idiosyncratic and adversarial error and reliably fails against common-mode error, and the two camps are measuring different failure distributions. The evidence further indicates that independence is partially recoverable, but along axes the field mostly does not vary - different model family, different evidence channel, non-neural verifier, and, most cheaply and most neglected, a protocol that denies evaluators a shared context before they commit. This paper reports no experiments and no measurements of its own. It states what the published record establishes, states flatly what it does not, proposes an instrument for the missing quantity, and names five experiments that would settle the open part.

1. Introduction

The negative result that motivates peer verification is well established. A model asked to critique its own candidate solution degrades rather than improves it on reasoning and planning tasks, while the same pipeline given a sound external verifier improves substantially, and merely re-prompting with that sound verifier retains most of the benefit of more elaborate architectures (Stechly et al., 2024). The verifier role, filled by the generating model itself, produces enough false positives to compromise the system that depends on it (Valmeekam et al., 2023). A companion analysis of single-model hallucination detection reached the same structural conclusion from a different corpus: the detection signals with demonstrated traction are all external to anything the model reports about itself (Mahendrakar, 2026). The field's response has been consistent and, on its face, reasonable. If the checker must be external, make the checker another model.

That move is now infrastructure. Agent benchmarks adopt a built-in judge and report numbers through it (Wang et al., 2026). Multi-agent failure taxonomies are annotated at scale by a judge pipeline (Cemri et al., 2025). Safety evaluations route verdicts through judge panels (Hu and Qu, 2026). Reward signals for on-policy reinforcement learning are judge outputs (Wang et al., 2026). The practice has a survey of its own, organised around the question of how to build a reliable judge (Gu et al., 2025), and the foundational result that licensed it - strong judges matching human preference agreement at rates comparable to human-human agreement (Zheng et al., 2023) - is among the most cited in the area.

The justification for the move, however, is not that a second model is accurate. It is that a second model is *different*. Verification is worth doing because the verifier might catch what the generator missed, and that is a claim about the relationship between two error distributions, not about either one alone. The claim has a name in the measurement literature - conditional independence - and it has an operational consequence: if the verifier's mistakes fall on the same items as the generator's, then a verdict of "looks fine" carries no information beyond what the generator already asserted, no matter how accurate the verifier is on average. Nothing in a reported accuracy, F1 or agreement figure distinguishes the two cases.

Three things now make this worth stating rather than assuming. First, the independence assumption has been tested directly, in settings adjacent to agent verification, and it fails at a magnitude that changes conclusions: a panel of nine frontier judges drawn from seven model families has been measured to carry roughly two independent votes' worth of information, with the panel's accuracy falling well short of what independent voting would deliver and the best single judge matching or beating the full panel (Kohli, 2026). Second, model mistakes appear to be growing *more* similar as capability increases, and judge scores favour models similar to the judge (Goel et al., 2025) - so the trend is against the assumption at exactly the point where human oversight becomes least available. Third, the verification error budget in agent settings is now large enough to move published conclusions: LLM judges misclassify 9 to 24 percent of mobile-agent trajectories, and judge accuracy varying from 76 to 91 percent shifts agent rankings by up to 13 positions across 24 agents (Wang et al., 2026), while an audit of 496 expert-reviewed tasks across four tool-calling benchmark families found 92 evaluator-human disagreements, an 18.5 percent misalignment rate, and concluded that current tool-calling scores can reflect evaluator artefacts rather than agent capability (Bhat et al., 2026).

There is a second audience for whom the argument runs differently. A reader who already knows that judges are noisy will say the problem is calibration and aggregation, and that the fix is a bigger panel with a better estimator. That reader has good evidence: panels of smaller models from disjoint families outperform a single large judge while costing far less (Verga et al., 2024), a robust aggregator dominates plain consensus under contamination (Acharya et al., 2026), and reweighting a verifier ensemble by a measured dependence structure beats majority voting (Kuai et al., 2026). This paper does not dispute any of that. It argues that those results and the pessimistic ones are answers to different questions, and that the difference is a difference in the assumed failure distribution rather than in the facts. Section 6: reconciliation sets this out.

The question this paper takes up is therefore not whether cross-agent verification works. It is:

Does adding a peer evaluator supply an error signal that is conditionally independent of the system being evaluated, or does it relocate a shared blind spot; is independence obtainable architecturally - by a different model family, a different evidence channel, a non-neural verifier, or a protocol that denies evaluators a shared context - rather than by adding an instance; and what would have to be measured and reported for a reader to tell the difference?

Four contributions follow. Section 2 stipulates the vocabulary the argument needs and states the one distinction the literature runs together. Section 3 records what cross-agent verification demonstrably buys, because the pessimistic reading is easy to overstate. Section 4 sets out four mechanisms by which the independence assumption fails and what each is and is not established to show. Section 5 asks where independence is partially recoverable, and finds the protocol-level answer better supported and cheaper than the architectural ones. Section 6 reconciles the two camps. Section 7 names the quantity nobody reports, notes that instruments to measure it already exist outside the agent literature, and proposes five experiments. Sections 8 and 9 state what would falsify this reading and where the evidence base itself is weak.

No experiments were run for this paper and no measurement in it is the author's. Every number is attributed to the work that produced it, and Section 9 is explicit that several load-bearing numbers come from single-author, unrefereed 2026 preprints.

2. What "independent" has to mean

Five terms carry the argument, and each is used here more narrowly than in common usage.

Peer verification denotes any arrangement in which the output, trajectory or intermediate reasoning of one language-model system is assessed for correctness, safety or task completion by another language-model system. This covers LLM-as-a-judge scoring, cross-agent critique, judge panels, verifier ensembles and the evaluator half of an actor-verifier loop. It excludes self-critique by the same model in the same context. That excluded case is the subject of the predecessor record to this one, which surveys single-model hallucination self-detection and concludes that the signals with demonstrated traction sit outside the introspective loop (Mahendrakar, 2026); the present paper begins where that one ends, by asking what happens when the external checker is itself a language model. It also excludes human evaluation.

Instance multiplicity we define as the number of distinct evaluator invocations contributing to a verdict. *Signal independence* we define as the degree to which an evaluator's errors are conditionally independent of the evaluated system's errors, given the item. The distinction is the whole argument. Instance multiplicity is what practitioners add and what papers report; signal independence is what the justification for peer verification requires and what almost nobody reports. They are not the same quantity and they are not monotonically related: a panel of nine can carry the information of two (Kohli, 2026), and seventeen vision-language models from eight families can behave as between two and a half and four independent voters (Bugaud, 2026).

Common-mode error we define as error that lands on the same items across evaluators for a reason the evaluators share - a shared pretraining corpus, a shared alignment recipe, a shared distillation ancestor, a shared prompt, a shared surface heuristic, or a shared context supplied at evaluation time. *Idiosyncratic error* is the remainder. The distinction matters because averaging is a variance reduction operator: it attacks idiosyncratic error and is inert against common-mode error. This is not a novel observation about ensembles; it is the standard one, and it has been given an information-theoretic form for LLM ensembles specifically, in which correlated errors modelled by a Gaussian copula impose a performance floor that additional models cannot cross (Turkmen et al., 2026). What is missing is not the theory. It is any measurement of where deployed cross-agent verification sits between the two regimes.

One consequence of the definitions should be stated before it is used, because it is the reason accuracy figures cannot settle the question. Two verifiers with identical accuracy can carry entirely different amounts of information about a generator's errors, depending on where those errors fall. A verifier whose mistakes coincide exactly with the generator's is worthless as a check however accurate it is, and a less accurate verifier whose mistakes fall elsewhere can be worth a great deal. Every reporting convention in current use - accuracy, balanced accuracy, F1, Cohen's kappa against human labels, agreement with a reference judge - is computed on the verifier alone. None of them is a function of the joint distribution. A reader of a benchmark table therefore cannot tell, even in principle, whether the verifier is supplying an independent signal.

Finally, a scope statement, placed here rather than at the end. This paper is about *cross-agent* detection of unreliability. It is not about failure attribution within a multi-agent trajectory - identifying which agent and which step caused a failure - which is a distinct task with its own benchmarks and its own poor results (Zhang, S., et al., 2025; Qiao et al., 2026), and it is not about cascade containment. It is also not about whether human agreement is the right criterion for validating an automatic judge, which

is a separate and live methodological dispute (Norman et al., 2026). Where those literatures supply evidence about the joint error structure of evaluators, that evidence is used; where they concern the validity of the evaluation construct itself, it is left alone.

3. What cross-agent verification demonstrably buys

The pessimistic reading of this literature is easy to overstate, so the positive results come first, and they are substantial.

Cross-agent judging is far better than the automated alternative it replaced. In an audit of 14,750 execution traces from thirteen tool-using agents across four domains, substring-heuristic judging of agent answers - the convention in a great many agent benchmarks - agreed with human annotation at chance, with Cohen's kappa of 0.049 and 0.015 against the two annotators, while single LLM judges on the same traces reached 0.336 to 0.567 and human annotators agreed with each other at 0.835 (Gurram, 2026). On injected traces the heuristic over-counted correctness by 34 percentage points, enough to reverse the direction of a baseline-versus-perturbed comparison. Whatever else is true, replacing string matching with a model judge was a large improvement.

Strong judges track human preference at rates comparable to human-human agreement, on the task they were validated for. Zheng et al. (2023) report over 80 percent agreement between GPT-4 judges and both controlled and crowdsourced human preferences on MT-Bench, the same level as agreement between humans, and this result is the licence under which most of the current practice operates. Section 4.2: similarity as bias records a specific and serious complication to how that agreement should be read, but the finding itself is not in dispute.

Purpose-trained verifiers can be small, cheap and accurate. A 770M-parameter fact-checking model reaches GPT-4 accuracy on grounded fact-checking at roughly 400 times lower cost (Tang et al., 2024), which is the clearest existence proof that verification quality and verification cost are separable. Training verifiers with a next-token objective jointly on verification and generation yields large best-of-N gains over discriminative verifiers, DPO verifiers and LLM-as-a-judge (Zhang, L., et al., 2025). A verifier fine-tuned specifically for trajectory-level failure verification improves attribution over prior methods across open-source and API backbones (Qiao et al., 2026). Cross-agent verification is not a fixed capability of general models; it responds to training.

Panels can beat single judges, and robust aggregation can beat plain panels. A panel of several smaller models drawn from disjoint families outperforms a single large judge across three judge settings and six datasets, exhibits less intra-model bias, and costs over seven times less (Verga et al., 2024). Framing panel consensus as robust mean estimation, replacing the mean with a geometric median dominates plain consensus on every biased corruption type tested across thirteen open-weight judges, and a three-judge robust committee at 38B parameters beats a 675B single model under 30 percent bimodal-random corruption (Acharya et al., 2026). Reweighting a verifier ensemble by a measured dependence structure yields 3.5 and 2.6 percentage-point gains in accuracy and precision over majority voting (Kuai et al., 2026).

Structured cross-model protocols help non-expert judges reach truth. In a setting where debating experts hold information the judge lacks, debate raises non-expert model accuracy to 76 percent and non-expert human accuracy to 88 percent, against naive baselines of 48 and 60 percent, and optimising debaters for persuasiveness improves the non-expert's ability to identify truth (Khan et al., 2024). Extending the comparison across mathematics, coding, logic and multimodal asymmetries, debate outperforms consultancy on all tasks tested when the consultant is randomly assigned which answer to argue for, and stronger debater models increase judge accuracy, though more modestly than in earlier studies (Kenton et al., 2024). Those authors are careful about the comparison that matters more: against

direct question answering with no debate at all, debate wins on extractive QA with information asymmetry and the results are mixed on tasks without it. Training a prover against a small verifier increases both the prover's accuracy and the verifier's robustness to adversarial attack, and the resulting legibility transfers to time-constrained humans (Kirchner et al., 2024).

Taken together, these results establish that cross-agent verification is a real capability with real headroom, and that a blanket claim that peer verification does not work would be false. The rest of this paper is about what those results do not establish: none of them measures whether the verification signal is independent of the generator, and several of them, read closely, are evidence that it is not.

4. Four ways the independence assumption fails

4.1 Panels lose most of their nominal independence to shared mistakes

The most direct test is also the most recent. Kohli (2026) evaluates a panel of nine frontier judges drawn from seven model families on three natural-language inference datasets with 100 human annotations per item, and quantifies the panel's informational value using Kish effective sample size against a Condorcet null. The nine judges supply approximately two independent votes' worth of information; roughly three-quarters of nominal independence is lost because the models err on the same items. The consequences reported are that the panel's accuracy falls 8 to 22 percentage points short of what independent voting would achieve, that the best single judge matches or outperforms the full panel across all conditions, and that established aggregation methods close at most 11 percent of the gap even when given access to correct answers. The deficit is reported to hold up when prompts are varied, when temperature is changed, when chain-of-thought reasoning is used, and on a pairwise preference task.

The scope of that result should be stated precisely, because it is narrower than its framing suggests. It concerns natural-language inference and a preference task, not agent trajectories; it is a single-author preprint; and "the best single judge matches or outperforms the panel" is a statement about post-hoc identification of the best judge, which a deployer choosing a panel in advance cannot perform. What the result does establish, and what matters here, is that the gap between instance multiplicity and signal independence is not a rounding error in a setting where seven distinct families were deliberately used.

Two independent measurements point the same way. In vision-language ensembles, 17 models from 8 families collapse to an effective dimensionality of 2.5 to 3.6 independent voters, and correlated majority errors create a tier of items - between 1.5 and 6.5 percent of questions - on which ensemble accuracy is driven to zero despite the best member being correct (Bugaud, 2026). In the tool-agent audit discussed above, a three-model majority-vote ensemble reached kappa 0.432 against human labels while its strongest member alone reached 0.567; the ensemble was conservatively biased, marking 25 percent of traces correct against humans' 38 percent, with disagreements dominated by rejecting answers humans accepted (Gurram, 2026). Two qualifications are load-bearing here and the paper's own text supplies both: the ensemble-versus-best-member difference sits within overlapping bootstrap confidence intervals on a 100-label calibration set, and two of the three ensemble members were from the same provider. The second qualification is not a defect of the finding but an instance of the phenomenon under discussion.

4.2 Similarity is not merely redundancy; it is a scoring bias

Correlated errors would be a smaller problem if they were symmetric noise. They are not. Goel et al. (2025) propose Chance Adjusted Probabilistic Agreement, a similarity metric defined on overlap in model mistakes, and report that LLM-as-a-judge scores favour models similar to the judge, generalising

earlier self-preference results. That earlier result is itself mechanistic rather than incidental: LLM evaluators can distinguish their own outputs from others' at non-trivial accuracy, and fine-tuning reveals a linear correlation between self-recognition capability and self-preference bias that survives controlled confound tests (Panickssery et al., 2024). Goel et al. further report that model mistakes are becoming more similar as capabilities increase, which makes the problem prospective rather than historical.

A statistical framework for the same phenomenon has since been developed. Kuai et al. (2026) audit behavioural entanglement among 18 black-box LLMs from six families, attributing it to shared pretraining data, distillation and alignment pipelines, and report that their dependence indices are associated with judge over-endorsement bias on a disjoint evaluation set, with Spearman correlations of 0.508 and 0.520 ($p < 0.01$) on MMLU-Pro and 0.441 and 0.457 ($p < 0.05$) transferring to MATH-500. These are associations between a similarity index and a bias, not demonstrations that similarity causes the bias, and the authors' own framing is that apparent agreement can reflect shared error modes rather than independent validation.

There is also a complication to the foundational agreement result that belongs here. Norman et al. (2026), evaluating 21 judges from nine providers over approximately 541,000 judgments, report that exact-match agreement - the metric in routine use for judge validation - does not correct for chance and systematically overstates discriminative ability, with deflation between exact match and Cohen's kappa of 33 to 41 percentage points on MT-Bench, and judge rankings shifting by up to 14 positions across benchmarks. Their finding does not contradict Zheng et al. (2023); it changes what the 80 percent figure licenses. High raw agreement between a judge and a human is consistent with both tracking the same surface cue. That possibility is the human-side analogue of the machine-side correlation this section is about, and it is the subject of an active methodological dispute that this paper deliberately does not adjudicate.

4.3 Shared inputs destroy independence that the roster appears to supply

Before the social mechanism, a mechanical one. Panel members are almost always given the same rubric, and in the tool-agent audit the three ensemble members received a byte-identical instruction (Gurram, 2026). A shared instruction is a shared input, and evaluator verdicts are sensitive to it in ways that have nothing to do with the behaviour being evaluated. Weng et al. (2026), stress-testing four agent-class safety judges on trajectories drawn from ASSEBench and R-Judge, report that rewording an evaluation policy without changing what it says flips up to 9.1 percent of verdicts beyond baseline jitter, that 18 to 43 percent of all the flips they observe under such rewrites fall on cases that are not ambiguous, and that a judge moves about as much when the policy's norms really change as when only its surface structure is reshuffled, with no ability to distinguish the two. They conclude that current safety scores mix up the agent's behaviour with the wording of the evaluator's prompt; the conclusion this paper draws in addition is that a rubric-induced verdict shift is common-mode by construction, because every panel member receives the same rubric. Consistent with this, a reliability harness run with four state-of-the-art judges on four benchmarks, which between them cover safety, persuasion, misuse and agentic behaviour, finds no judge uniformly reliable across benchmarks, and its preliminary experiments expose consistency problems when model-produced responses receive simple formatting edits, are paraphrased, are made more or less verbose, or have their ground-truth label flipped (Dev et al., 2026). Both are diagnostic tools reporting variation rather than causal accounts, and neither measures generator-verifier overlap; what they establish is that a channel for common-mode error exists upstream of the models entirely.

The social mechanism is more striking, because independence can be removed at evaluation time from a panel that would otherwise have it. Hu and Qu (2026) run a controlled two-round experiment across

six open-weight LLMs and six datasets in which each model first judges an item alone and then judges it again after six simulated peers either assert a wrong label or abstain. Averaged over reviewers, the false-alarm rate is 56.5 percent when the peers stay silent but 87.5 percent when they assert the wrong label, and in that wrong-label condition a majority vote over the final judgments puts the panel's false-alarm rate at 100 percent. The effect is strongly asymmetric, with reviewers following pushes toward "unsafe" far more than pushes toward "safe", approximately 75 percent against 17 percent. The sentence that matters most for this paper is the control condition: without an asserted label, the same panel outperforms its average member. Independence, in that experiment, is a property of the protocol, not of the roster.

The same structure appears in multimodal panels under a deliberate attack. Shu (2026) reports that quoting peer visual judgments into a VLM panel creates anchoring gaps of 19 to 26 percentage points under both self and peer framing, while a matched-content, label-only control changes the outcome by -0.17 percentage points, indicating the effect is carried by the quoted content rather than by the provenance label. The proposed defence is precisely a restoration of protocol independence: cross-checking a quote against independently collected blind votes blocks 84.9 percent of fabricated attacks. The author is explicit that the comparison between deliberately generated wrong quotes and naturally occurring ones measures differential damage under a specific tested construction rather than a provenance-only causal effect.

Debate and deliberation research supplies the same lesson from the cooperative side. Across a controlled study of ten homogeneous agents over three rounds on GSM-Hard and MMLU-Hard, sycophantic conformity reaches modal adoption of up to 85.5 percent, peer rationales destabilise previously correct reasoning at vulnerability rates up to 70.0 percent, and plurality voting discards correct answers already present in the pool with an oracle gap up to 32.3 percentage points, while consuming 2.1 to 3.4 times the tokens of isolated self-correction for equal or lower accuracy (Bertalanic and Fortuna, 2026). That study is scoped to the 7-8B parameter class with unstructured roles and its authors say so. Decomposing stance changes in debate into distinct mechanisms, Hao et al. (2026) report that agents which otherwise hold their ground still take on a peer's error 20 to 39 percent of the time when the peer's reasoning has no content at all, that strict conformity does mostly harm (57 to 77 percent correct-to-wrong across model replications), and - the finding most relevant here - that making agents adopt peer answers less often does not raise accuracy unless correctness labels or self-reflection controls are available, since without them the influence that helps cannot be told apart from the influence that hurts. Debate can also lower accuracy over rounds even when stronger models outnumber weaker ones, with models shifting from correct to incorrect answers in response to peer reasoning (Wynn et al., 2025). And disentangling the components of multi-agent debate across seven benchmarks, majority voting alone accounts for most of the gains usually attributed to debate, with a theoretical account under which debate induces a martingale over belief trajectories and therefore does not by itself improve expected correctness (Choi et al., 2025).

4.4 Optimisation pressure degrades a peer verifier, faster on shared substrate

The fourth mechanism appears when the verifier stops being a measuring instrument and becomes a target. Optimising against an imperfect proxy reward degrades ground-truth performance, with a functional form that depends on whether the optimisation is reinforcement learning or best-of-n sampling and with coefficients that scale smoothly in the reward model's parameter count (Gao et al., 2022). In-context, the same effect arises spontaneously: iterative self-refinement against a model evaluator causes the evaluator's ratings to improve while human-judged quality stagnates or declines, and the two factors reported to affect severity are model size and *context sharing between the generator and the evaluator* (Pan et al., 2024). That second factor is this paper's variable, varied deliberately - and it is the only study in this corpus that does so.

The theoretical counterpart is sharper than the empirical one. Acharya et al. (2026) show that a single judge failing in a biased way typical of LLMs, such as mode collapse, sycophancy or safety refusal, suffices to give a panel reporting consensus scores unbounded bias at any positive contamination rate, no matter how large the jury is. Jury size is the wrong knob; the aggregator is the right one. Their robust estimator is a genuine fix for that failure class, and Section 6: reconciliation returns to why it is not a fix for the class in Section 4.1: correlated panels.

Finally, verification failure is not merely additive with generation failure. Jin et al. (2026), evaluating six open-weight vision-language models across five medical VQA datasets, report that because verifier and generator are capacity-coupled, the verifier can over-agree with the generator, and that a wrong answer from the generator comes with a *higher* probability of both verifier error and agreement bias. They also report that verifiers give image evidence less attention than generators do, and that in actor-verifier loops running over multiple turns, false verification locks in most of the answers that were wrong to begin with. Cross-verification, in their experiments, reduces but does not eliminate the effect. Two caveats belong with any citation of this work: the posted version carries an unresolved placeholder where its framework is named, which is a sign of an unfinished manuscript, and the domain is medical VQA rather than agent trajectories. What it contributes is the only direct measurement in this corpus of the conditional structure the whole argument turns on - verifier error correlated with generator error, on the same items.

5. Where independence is partially recoverable

If instance multiplicity is the wrong knob, the question is which knobs are the right ones. The evidence supports four candidate axes, unevenly.

Different model family: helps, insufficiently, and the amount is measurable. Every result in Section 4.1: correlated panels was obtained on rosters that already varied family - seven families in Kohli (2026), eight in Bugaud (2026), six in Kuai et al. (2026) - so family variation is demonstrably not sufficient. It is also demonstrably not nothing. Wang et al. (2026) report that different LLM backends produce qualitatively opposite failure profiles on mobile-agent trajectories: GPT-based judges produced 48 hard-core failures dominated by false negatives (30 against 18), behaving conservatively, while Qwen-based judges produced 78 dominated by false positives (71 against 7), behaving permissively. Two verifiers with opposite error asymmetries are genuinely complementary in a way two conservative verifiers are not. The same analysis, however, identifies surface UI match as a weakness shared across backends, which is the shape the whole argument predicts: family variation moves the idiosyncratic component and leaves a common-mode residue. Family-aware aggregation that votes within families before voting across them recovers 18 to 26 percentage points on the tier of items where correlated majority errors are decisive (Bugaud, 2026), which indicates that the correlation structure is not only measurable but exploitable.

Different evidence channel: promising, barely explored. In the mobile-agent judge study, a simple baseline judge given only sampled screenshots - no UI metadata and no agent reasoning - is competitive with and often exceeds purpose-built judge pipelines, and among competitive methods the backbone rather than the pipeline is the primary driver of quality (Wang et al., 2026). Read alongside the finding that verifiers under-attend to image evidence relative to generators (Jin et al., 2026), the natural hypothesis is that what a verifier is *shown* changes its error distribution more than what it is *told to do*. That hypothesis is not tested anywhere in this corpus. A verifier that sees only the environment state, and never the generator's reasoning, is denied the most obvious channel for common-mode error; whether that buys independence is an open empirical question, and experiment E2 in Section 7: what is not measured is designed to answer it.

Non-neural and purpose-trained verifiers: the strongest evidence, the weakest coverage. A sound

external verifier produces significant performance gains where self-critique produces collapse (Stechly et al., 2024). In the one training comparison available here, a rule-based programmatic checker used as the reward signal produced a 36.8 percent success rate against 30.2, 27.9 and 26.4 percent for three LLM-judge backends on the full 116-task suite (Wang et al., 2026). That comparison carries a caveat the authors state plainly and which must travel with the number: the rule-based reward coincides with the evaluation signal, so its result is best read as an upper bound rather than as a fair comparison. The cost of the approach is coverage, not accuracy - rule-based checkers require per-task engineering and do not extend to new tasks or apps - and rule-based evaluation used as ground truth has its own bias, tending to underreport web-agent success relative to expert review (Lu et al., 2025). Purpose-trained neural verifiers occupy the middle ground and are the best-evidenced compromise: a 770M model at GPT-4 accuracy and 400 times lower cost (Tang et al., 2024), and a trajectory-level verifier that improves attribution across backbones (Qiao et al., 2026).

Protocol independence: cheapest, best-evidenced, and mostly neglected. The single largest independence effect in this corpus is not architectural. It is the difference between evaluators that commit before seeing each other and evaluators that do not. The same panel that reaches a 100 percent false-alarm rate after hearing an asserted peer label outperforms its average member when peers stay silent (Hu and Qu, 2026). Cross-checking against independently collected blind votes blocks the large majority of a fabricated-quote attack (Shu, 2026). Debate helps specifically where the debaters hold information the judge lacks (Khan et al., 2024; Kenton et al., 2024), which is an engineered asymmetry rather than a roster property. Regulating which peers an agent may reference during debate, rather than scoring agents individually, is reported to mitigate blind conformity, and the same work notes that confidence- and perplexity-based individual evaluation can worsen it (Wu et al., 2026). Against this, unstructured homogeneous debate is worse than isolated self-correction at 2.1 to 3.4 times the token cost (Bertalanic and Fortuna, 2026). The generalisation the evidence supports is narrow but clear: blind, committed evaluation preserves whatever independence the roster supplies, and shared context consumes it.

6. What the two camps are actually disagreeing about

The literature reads as contradictory. Panels beat single judges (Verga et al., 2024); panels lose to their best single member (Kohli, 2026; Gurram, 2026). Better aggregation fixes panels (Acharya et al., 2026; Kuai et al., 2026); no aggregator closes more than a small fraction of the gap (Kohli, 2026). Debate helps (Khan et al., 2024; Kenton et al., 2024); debate is a martingale and adds nothing over voting (Choi et al., 2025).

The claims are compatible, and the vocabulary of Section 2: what independence has to mean is what makes them so. Aggregation is a variance-reduction operator over an assumed error distribution, and the three camps assume different ones.

Idiosyncratic error, meaning independent noise across evaluators, is the regime averaging was built for: more instances help at the usual rate. This is where panel-beats-single-judge results are obtained, and it is also why multi-trial aggregation is a sound recommendation for a single noisy judge. Repeated identical evaluations flip pairwise preferences 13.6 percent of the time on average, 28 percent of questions exceed a 20 percent flip rate, and eleven trials are needed for a majority vote to recover a 50-trial reference verdict with 95 percent probability (Yagubyan, 2026).

Contaminated-minority error, meaning a bounded fraction of evaluators failing in a biased, arbitrary or adversarial way, breaks the mean but not aggregation itself: the mean is the wrong statistic and a robust estimator is the right one. This is exactly the regime Acharya et al. (2026) formalise, and in which their geometric-median aggregator dominates plain consensus.

Common-mode error, meaning the majority erring together for a shared reason, is the regime in which neither more instances nor a better estimator helps, because the quantity being estimated is itself biased. This is the regime Kohli (2026) measures, Turkmen et al. (2026) give an information-theoretic floor for, and Hu and Qu (2026) induce experimentally by supplying a shared cue.

Read this way, the disagreement is not about whether panels work. It is an unstated disagreement about which regime deployed cross-agent verification is in, and for agent trajectories that regime has not been measured. That is a gap rather than a controversy: no paper in the corpus surveyed here reports, for an agent-verification setting, the conditional dependence between verifier error and generator error.

Two further observations follow from the same framing. First, the choice of aggregator is a *declaration* about the assumed failure regime, and papers should say which one they are assuming; a robust estimator deployed against common-mode error is not merely ineffective but misleadingly reassuring. Second, calibration budgets should be spent accordingly - the practical question in fitting a panel calibrator has been posed precisely as whether the next judge's information is estimable under the available human labels (Zhu et al., 2026), which is the finite-sample form of the same question.

7. What is not measured, and five experiments that would settle it

The instrument this field needs mostly already exists; it has simply never been pointed at agent verification. Chance Adjusted Probabilistic Agreement is defined on overlap in model mistakes and was introduced precisely so that similarity could be reported and corrected for (Goel et al., 2025). Kish effective sample size against a Condorcet null converts a panel roster into a number of effective votes (Kohli, 2026). Behavioural entanglement indices characterise the joint failure manifold of black-box models (Kuai et al., 2026). Effective ensemble dimensionality and family-aware aggregation do the same for vision-language ensembles (Bugaud, 2026). What is missing is the reporting convention, and the normative claim of this paper is a single sentence:

A paper that reports a verifier's accuracy against ground truth should also report that verifier's error overlap with the system it verifies.

That is one additional column in a table the authors have already computed the inputs for. Five experiments would convert the recommendation into knowledge.

E1 - Measure the common-mode fraction for agent trajectories. Take a trajectory corpus with expert per-trajectory labels and multiple agent models - AgentRewardBench supplies 1,302 trajectories across five benchmarks and four LLMs with expert review (Lu et al., 2025), and MobileJudgeBench supplies 931 trajectories with 88.4 percent pairwise annotator agreement across six benchmarks, four agent models and 68 apps (Wang et al., 2026). For every generator-verifier pair, compute verifier error conditional on generator error and conditional on generator success, and report the ratio. Report Kish effective sample size for every panel. The deliverable is a single number per setting: what fraction of verification error is common-mode. Nothing about the design is expensive, because both corpora and both label sets already exist.

E2 - Test whether the evidence channel, not the model, carries independence. Hold the verifier backbone fixed and vary only what it is shown: final answer only; environment state only; full trajectory including the generator's reasoning. The prediction implied by the screenshot-only baseline result (Wang et al., 2026) and by the lazy-verifier finding (Jin et al., 2026) is that verifiers denied the generator's reasoning will have lower error overlap with the generator at equal or slightly lower marginal accuracy. If that prediction fails - if channel restriction costs accuracy without buying independence - the architectural route is in more trouble than this paper suggests.

E3 - Separate roster independence from protocol independence. The peers-assert-versus-peers-

abstain manipulation (Hu and Qu, 2026) has been run on safety classification with simulated peers. Run it on agent-trajectory judging with real peers, crossed with roster composition: same-family blind, cross-family blind, same-family shared-context, cross-family shared-context. The hypothesis the evidence supports is that the protocol factor is larger than the roster factor. If it is, the cheapest available intervention in cross-agent verification is a change to evaluation order, not to the model list.

E4 - Measure how fast a peer verifier degrades under optimisation pressure. Context sharing between generator and evaluator has been identified as a factor in reward-hacking severity (Pan et al., 2024) and proxy overoptimisation has a measured scaling form (Gao et al., 2022), but neither has been characterised for agent verification. Fix a task distribution, optimise a generator against a peer verifier, and track the divergence between verifier score and held-out ground truth as a function of verifier-generator error overlap measured at the start. The prediction is that initial overlap predicts the rate of divergence. That would make error overlap not merely a reporting nicety but a deployment-time safety parameter.

E5 - Establish how much reported agent progress is verifier artefact. An audit of four tool-calling benchmark families found 92 evaluator-human disagreements across 496 expert-reviewed tasks, an 18.5 percent misalignment rate, with 23 repeated evaluations of one identical setup producing scores from 57.9 to 76.8 percent, a spread the authors state is large enough to change leaderboard conclusions (Bhat et al., 2026). Independently, judge accuracy varying from 76 to 91 percent shifts agent rankings by up to 13 positions across 24 agents (Wang et al., 2026). Neither study re-scores a historical leaderboard. Doing so - re-running a published agent leaderboard under several validated judges and reporting rank stability - would convert a methodological worry into a quantity the field can argue about. Aggregate-score rankings are already known not to transfer out of distribution, and the authors of that argument state that existing evidence partly supports it but is too thin to confirm (Patel et al., 2026); E5 is a direct way to thicken it.

8. What would falsify this reading

Four findings would substantially undermine the position taken here.

First, a measurement showing that in agent-trajectory verification specifically - as opposed to natural-language inference, VQA and preference tasks - verifier error is close to conditionally independent of generator error. The results in Section 4.1: correlated panels would then be domain-specific rather than general, and instance multiplicity would be a reasonable proxy for independence after all. Nothing in this paper rules this out; the measurement has not been made.

Second, an aggregator that closes a large share of the independence gap without access to ground-truth labels. Kohli (2026) reports that established methods close at most 11 percent of the gap even with access to correct answers, but that is a statement about the methods tested, not an impossibility result, and dependence-aware reweighting already recovers a few points over majority voting (Kuai et al., 2026).

Third, evidence that model mistakes stop converging as capability increases, or diverge again. The prospective force of the argument rests on the observed trend toward more similar mistakes (Goel et al., 2025), and that trend is an empirical observation over a model set, not a law.

Fourth, a demonstration that the protocol effects in Section 4.3: shared inputs are artefacts of simulated peers and small open-weight models. The strongest protocol result uses six simulated peers and six open-weight models (Hu and Qu, 2026), and the strongest homogeneous-debate result is scoped to the 7-8B class (Bertalanic and Fortuna, 2026). If both fail to replicate on frontier models with genuine peers, the protocol-first recommendation in Section 5: partial recovery loses most of its support.

9. Limitations of this evidence base

The most serious limitation is provenance, and it should be the first thing a sceptical reader is told. Several load-bearing results here are single-author, unrefereed 2026 arXiv preprints: the effective-sample-size result (Kohli, 2026), the tool-agent judging audit (Gurram, 2026), the ensemble-dimensionality result (Bugaud, 2026), the run-to-run reliability study (Yagubyan, 2026), and the panel-anchoring attack (Shu, 2026). Three of those five carry numbers this paper leans on. One of the cited manuscripts contains an unresolved placeholder where its own framework is named (Jin et al., 2026). None has been replicated by an independent group, and the convergence between them is therefore weaker evidence than the count of papers suggests. A reader who discounts unrefereed single-author work should discount Section 4.1: correlated panels and Section 6: reconciliation accordingly. The venue-confirmed portion of the corpus is smaller than the reference count suggests: by the venue or DOI recorded in their own metadata, that portion is Zheng et al. (2023, NeurIPS Datasets and Benchmarks), Tang et al. (2024, EMNLP), Zhang, L., et al. (2025, ICLR), Wynn et al. (2025, ICML workshop), Choi et al. (2025, NeurIPS Spotlight), Dev et al. (2026, ICLR workshop), Bertalanic and Fortuna (2026, ACM) and Zhu et al. (2026, WISE). Everything else cited here, including Goel et al. (2025), Panickssery et al. (2024), Verga et al. (2024), Kenton et al. (2024) and Stechly et al. (2024), is cited from an arXiv preprint whose metadata names no venue; several of those are widely cited and some have since appeared at venues the arXiv record does not list, but this paper does not assert a venue it could not verify.

Second, domain transfer is assumed rather than demonstrated. The independence measurements come from natural-language inference (Kohli, 2026), visual question answering (Bugaud, 2026), medical VQA (Jin et al., 2026) and safety classification (Hu and Qu, 2026). The conclusions are drawn about agent verification. The argument for transfer is mechanistic - shared pretraining, distillation and alignment are properties of the models rather than of the task (Kuai et al., 2026) - but a mechanism is not a measurement, which is why E1 is first in Section 7: what is not measured.

Third, this paper conducts no meta-analysis. The results are read and reported, not pooled, weighted or checked for publication bias. Where two studies conflict, Section 6 offers a reconciliation by regime; that reconciliation is an interpretation the author finds parsimonious, not a demonstrated fact, and an alternative reading in which the results simply disagree is available.

Fourth, the corpus is recent and probably incomplete. Thirty of the 41 references here are from 2025 and 2026, in an area publishing faster than it is being reviewed, and the search that assembled it was targeted rather than systematic. This is not a survey and does not claim coverage.

Fifth, a scope limit worth stating plainly: nothing here bears on whether a human verifier's errors are independent of a model generator's. That is the question scalable oversight actually needs answered, and the corpus surveyed here does not answer it.

10. Conclusion

The move from self-verification to peer verification was well motivated, and cross-agent checking is measurably better than the heuristics it replaced. What the move did not do is establish that a second model supplies an independent signal, and the evidence assembled here indicates that in several adjacent settings it largely does not: nominally diverse panels lose most of their independence to shared mistakes, similarity is itself a scoring bias, a shared evaluation context can remove independence that the roster appeared to supply, and optimisation pressure degrades a peer verifier faster when generator and evaluator share substrate. The apparent disagreement in the panel literature dissolves into a disagreement about which failure regime obtains, and in the corpus surveyed here that regime has not been measured for agent trajectories.

Independence is not a property of how many models are in the room. On the evidence available it is a property of what they are shown, what they were trained from, and whether they commit before they hear each other. All three are measurable with instruments that already exist, and none is currently reported. The one-column change proposed in Section 7: what is not measured - report error overlap with the system being verified, not only accuracy against ground truth - is the smallest intervention that would let a reader tell the difference.

Acknowledgements

This paper cites no unpublished communications and reports no funding. The author is solely responsible for the selection of the literature, which was targeted rather than systematic, and for the interpretive framing in Sections 2 and 6, which is not a claim made by any cited work.

Revision note

Version 1.1 (29 September 2026) replaces version 1.0. Passages that followed the wording of cited abstracts too closely are now either quoted, with their citation, or rewritten in new words, and a few qualifiers the sources state (such as "may" and "often") were restored in the process. The argument, the references and the reported values are unchanged.

References

- Acharya, A., Pan, K. W., & Verkhovsky, B. (2026). RoPoLL: Robust Panel of LLM Judges. arXiv:2606.30931. [PREPRINT]
- Bertalanic, B., & Fortuna, C. (2026). The Cost of Consensus: Isolated Self-Correction Prevails Over Unguided Homogeneous Multi-Agent Debate. ACM Conference on AI and Agentic Systems. doi:10.1145/3786335.3813137; arXiv:2605.00914.
- Bhat, V., Vaghasiya, J., Mohsin, M. A., & Aali, A. (2026). Benchmarking the Benchmarks: A Validity Audit of Tool-Calling Evaluation. arXiv:2607.02577. [PREPRINT]
- Bugaud, Z. (2026). Hidden Clones: Exposing and Fixing Family Bias in Vision-Language Model Ensembles. arXiv:2603.17111. [PREPRINT - single author, unrefereed]
- Cemri, M., et al. (2025). Why Do Multi-Agent LLM Systems Fail? arXiv:2503.13657. [PREPRINT]
- Choi, H. K., Zhu, X., & Li, S. (2025). Debate or Vote: Which Yields Better Decisions in Multi-Agent Large Language Models? NeurIPS 2025 (Spotlight). arXiv:2508.17536.
- Dev, S., Sloan, A., Kavner, J., Kong, N., & Sandler, M. (2026). Judge Reliability Harness: Stress Testing the Reliability of LLM Judges. Agents in the Wild: Safety, Security, and Beyond Workshop at ICLR 2026. arXiv:2603.05399.
- Gao, L., Schulman, J., & Hilton, J. (2022). Scaling Laws for Reward Model Overoptimization. arXiv:2210.10760. [PREPRINT]
- Goel, S., et al. (2025). Great Models Think Alike and this Undermines AI Oversight. arXiv:2502.04313. [PREPRINT]
- Gu, J., et al. (2025). A Survey on LLM-as-a-Judge. arXiv:2411.15594. [PREPRINT]
- Gurram, B. (2026). Auditing Automated Evaluation, Error Propagation, and Runtime Mitigation in Tool-Using Language Agents. arXiv:2604.16706. [PREPRINT - single author, unrefereed]
- Hao, X., et al. (2026). Not All Flips Are Conformity: Decomposing Stance Convergence in Multi-Agent LLM Debate. arXiv:2606.00820. [PREPRINT]
- Hu, Y., & Qu, J. (2026). Social Pressure Breaks Majority Voting in LLM Safety Panels. arXiv:2608.04415. [PREPRINT]
- Jin, R., Zhao, B., Kang, M., Zhang, Q., & Li, X. (2026). Verification Mirage: Mapping the Reliability Boundary of Self-Verification in Medical VQA. arXiv:2605.10850. [PREPRINT - posted version carries an unresolved placeholder for its framework name]
- Kenton, Z., et al. (2024). On scalable oversight with weak LLMs judging strong LLMs. arXiv:2407.04622. [PREPRINT]
- Khan, A., et al. (2024). Debating with More Persuasive LLMs Leads to More Truthful Answers. arXiv:2402.06782.

[PREPRINT]

Kirchner, J. H., Chen, Y., Edwards, H., Leike, J., McAleese, N., & Burda, Y. (2024). Prover-Verifier Games improve legibility of LLM outputs. arXiv:2407.13692. [PREPRINT]

Kohli, G. (2026). Nine Judges, Two Effective Votes: Correlated Errors Undermine LLM Evaluation Panels. arXiv:2605.29800. [PREPRINT - single author, unrefereed]

Kuai, C., et al. (2026). A Statistical Framework for Auditing Behavioral Dependence and Induced Bias in LLM Judges. arXiv:2604.07650. [PREPRINT]

Lu, X. H., et al. (2025). AgentRewardBench: Evaluating Automatic Evaluations of Web Agent Trajectories. arXiv:2504.08942. [PREPRINT]

Mahendrakar, P. M. (2026). The Self-Verification Gap: Why Working Hallucination Detectors Are External, and What Real-Time Correction Would Actually Require. Zenodo. doi:10.5281/zenodo.22030500.

Norman, J. D., Rivera, M. U., & Hughes, D. A. (2026). Reliability without Validity: A Systematic, Large-Scale Evaluation of LLM-as-a-Judge Models Across Agreement, Consistency, and Bias. arXiv:2606.19544. [PREPRINT]

Pan, J., He, H., Bowman, S. R., & Feng, S. (2024). Spontaneous Reward Hacking in Iterative Self-Refinement. arXiv:2407.04549. [PREPRINT]

Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. arXiv:2404.13076. [PREPRINT]

Patel, D. C., et al. (2026). Beyond Static Leaderboards: Predictive Validity for the Evaluation of LLM Agents. arXiv:2606.19704. [PREPRINT]

Qiao, H., Tong, H., Lim, E.-P., Liu, B., & Pang, G. (2026). VerifyMAS: Hypothesis Verification for Failure Attribution in LLM Multi-Agent Systems. arXiv:2605.17467. [PREPRINT]

Shu, Y. (2026). Forged Peer Judgments Mislead Multimodal LLM Judge Panels: Source-Blind Anchoring and Panel-Consensus Verification. arXiv:2608.07920. [PREPRINT - single author, unrefereed]

Stechly, K., Valmeekam, K., & Kambhampati, S. (2024). On the Self-Verification Limitations of Large Language Models on Reasoning and Planning Tasks. arXiv:2402.08115. [PREPRINT]

Tang, L., Laban, P., & Durrett, G. (2024). MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. EMNLP 2024. arXiv:2404.10774.

Turkmen, Y., Buyukates, B., & Bastopcu, M. (2026). Don't Always Pick the Highest-Performing Model: An Information Theoretic View of LLM Ensemble Selection. arXiv:2602.08003. [PREPRINT]

Valmeekam, K., Marquez, M., & Kambhampati, S. (2023). Can Large Language Models Really Improve by Self-critiquing Their Own Plans? arXiv:2310.08118. [PREPRINT]

Verga, P., et al. (2024). Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. arXiv:2404.18796. [PREPRINT]

Wang, Z., et al. (2026). Benchmarking LLM Judges for Mobile Agent Evaluation. arXiv:2608.11434. [PREPRINT]

Weng, S., Feng, Y., & Xie, X. (2026). Beyond Accuracy: Policy Invariance as a Reliability Test for LLM Safety Judges. arXiv:2605.06161. [PREPRINT]

Wu, H., et al. (2026). Group Perspective Matters: Regulating Debate Relationships Can Mitigate Blind Conformity in Multi-Agent Debate. arXiv:2608.03648. [PREPRINT]

Wynn, A., Satija, H., & Hadfield, G. (2025). Talk Isn't Always Cheap: Understanding Failure Modes in Multi-Agent Debate. ICML MAS Workshop 2025. arXiv:2509.05396.

Yagubyan, A. (2026). The Coin Flip Judge? Reliability and Bias in LLM-as-a-Judge Evaluation. arXiv:2606.13685. [PREPRINT - single author, unrefereed]

Zhang, L., Hosseini, A., Bansal, H., Kazemi, M., Kumar, A., & Agarwal, R. (2025). Generative Verifiers: Reward Modeling as Next-Token Prediction. ICLR 2025. arXiv:2408.15240. (Cited in text as Zhang, L., et al., 2025.)

Zhang, S., et al. (2025). Which Agent Causes Task Failures and When? On Automated Failure Attribution of LLM Multi-Agent Systems. arXiv:2505.00212. [PREPRINT] (Cited in text as Zhang, S., et al., 2025.)

Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmarks Track. arXiv:2306.05685.

Zhu, B., Xie, Y., & Rao, Y. (2026). A Finite-Calibration Regime Map for LLM Judge Panels. WISE 2026. arXiv:2606.01034.

(c) 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

The literature search, drafting and citation verification for this paper were carried out with AI assistance under the author's direction. Every citation was machine-verified against the arXiv API and Crossref before inclusion, and every quantitative claim was read back against the cited source's own text. The author is responsible for the final text and for all claims made in it.

DOI: 10.5281/zenodo.22044419