GRAIN: GRANULARITY AND THE VALIDITY GAP IN LLM-AGENT CONSENSUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Byzantine validity requires that a value proposed by every honest participant be decided, and it treats each honest proposal as a fixed input. A language-model agent's proposal is a random draw, and whether two draws count as the same proposal depends on the *grain* at which outputs are compared. The validity clause is therefore triggered only when all honest proposals collide, an event whose probability $A(h) = \sum_c p(c)^h$ falls as the grain gets finer; single-agent samples predict pre-interaction live-group agreement with no fitted parameter. We prove that, as groups grow, a sound observer certifies the honest mode with probability tending to one below a fault fraction of $\Delta/(2+\Delta)$ and to zero above it, where Δ is the honest top-two margin; the classical one third is the deterministic case. Under conditionally independent honest labels and a common view, the honest mode on three models is certified at the classical budget with mean probability only 0.84 even when only final answers are compared; on 24 annotated model–task cells, coverage at the human-judged grain of the stated decisive operation is about half its answer-grain value (0.43 versus 0.82; 95% interval 0.27–0.61). Measuring finer grains is itself unreliable: exact string matching splits 97% (weighted) of pairs that humans judge equivalent, while LLM judges reproduce plain answer equality. Prompted "Byzantine" agents, too, carry out their assigned attack to very different degrees across models. We close with a reporting checklist for consensus claims over stochastic agents.

1 INTRODUCTION

An ensemble of language-model agents can agree on an answer, agree on a short explanation, or agree to accept a particular output. These are different events. Self-consistency selects answers from repeated samples; debate lets agents revise proposals after communication (Wang et al., 2023; Du et al., 2024; Chen et al., 2024). In each case the meaning of an agreement rate depends on which differences between outputs the system chooses to retain.

Byzantine fault tolerance (BFT) supplies a language for such systems: *agreement*, *validity*, *termination* (Lamport et al., 1982; Castro & Liskov, 1999). Classical validity says that if every honest participant proposes v, then v is decided. The clause is stated over an execution in which each honest proposal is a fixed value, so "every honest participant proposes v" is a fact about that execution; randomised protocols keep this reading by placing randomness in coins, not in proposals (Ben-Or, 1983). A language model's proposal is instead a draw from a distribution, taken before the protocol runs, and whether two draws count as "the same proposal" is a design choice—the *grain* of agreement, written g_x below.

The transfer gap is not that LLM agents cannot run a BFT protocol. It is that the validity premise used to interpret the protocol changes from a fixed proposal condition to a random event whose probability depends on the agreement relation. For h honest agents with class law p, that probability is

$$A(h) = \sum_c p(c)^h,$$

and refining the partition cannot raise it. Where $A(h)$ is small the validity clause is satisfied vacuously (Beer et al., 2001; Kupferman & Vardi, 2003): a protocol may decide anything and remain correct. This does not make agreement infeasible—without faults, a simple protocol satisfies the clause on every execution while $A(h)$ is arbitrarily small—but it removes the clause as the thing that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

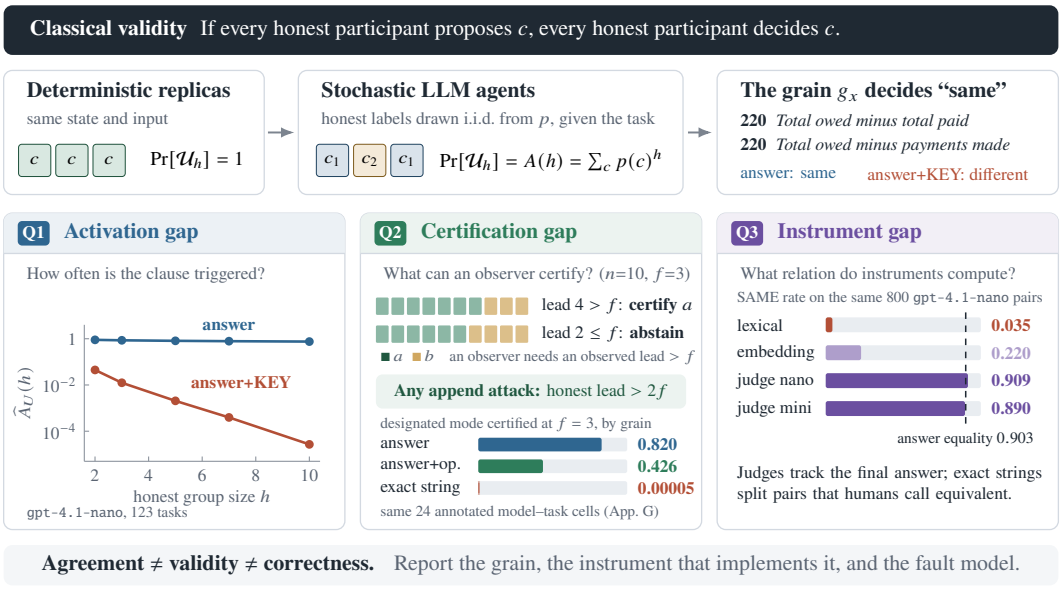

Figure 1: **The validity gap.** For deterministic replicas the validity antecedent $\mathcal{U}_h$ always holds; for stochastic agents it is a collision event with probability $A(h)$, and the grain g_x sets that probability. The gap appears in how often the clause is triggered (Q1, §6.1), what an observer can certify (Q2, §6.2) and which relation an instrument computes (Q3, §6.3). The grain example is a real pair of `gpt-4.1-nano` replies to one GSM8K task.

makes a group's answer trustworthy. We call this a validity *informativeness* gap: the clause is never violated; it stops constraining. Fig. 1 shows the mechanism and the three places where the gap can be measured.

Three gaps, three questions. **Activation gap (Q1).** How does the grain of agreement change how often classical validity is triggered? **Observability and certification gap (Q2).** When an observer sees all participants but the clause quantifies over honest ones, what can the observed histogram certify, and what fault budget survives? **Instrument gap (Q3).** Which relation do candidate agreement instruments actually compute, and how far is each from the intended one?

Contributions.

1. **Activation as a measurable quantity** (§4.1). Propositions 1–3 turn validity over stochastic proposals into $A(h)$, a quantity indexed by the grain.
2. **Certification limited by dispersion** (§4.2). Building on the observed-histogram criterion of Xu et al. (2026), we show that the honest count margin must exceed $2f$ for every rule sound for honest-plurality validity, and prove a sharp threshold: for fault fractions $\beta \neq \Delta/(2+\Delta)$, the honest mode is certified with probability tending to one exactly when β is below $\Delta/(2+\Delta)$, with Δ the top-two margin of the honest label law. The classical $1/3$ is the deterministic case $\Delta = 1$; a context shared by the agents makes the threshold itself random.
3. **Measurements of all three gaps** (§6). Single-agent samples predict live-group unanimity, homogeneous or mixed, with no fitted parameter. At $n = 10$, $f = 3$ the honest mode is certified with mean per-task probability only 0.843 even at the answer grain; on 24 annotated model–task cells, human partitions by stated decisive operation give 0.43 against 0.82 for answers on the same replies, and the 0.004 obtained with exact string matching (all 369 cells) is mostly an instrument artefact. Prompted attackers carry out 3.5–78% of their assigned attack in round 2, depending on the model. The lexical rule rejects 97% (weighted) of human-equivalent pairs and LLM judges track answer equality: no tested instrument measures the decisive-operation relation.

We do not claim that language-model groups cannot reach agreement, that the certificate bounds the fault tolerance of arbitrary protocols (it binds every sound rule, but only under one label law and a common view), or that the lexical grain measures reasoning.

2 Related Work

Consensus specifications. Byzantine agreement studies agreement, validity and termination under stated communication and fault assumptions (Lamport et al., 1982; Dwork et al., 1988; Castro & Liskov, 1999); authentication and synchrony change the achievable threshold (Dolev & Strong, 1983). Randomised protocols place randomness in the coins, not the inputs (Ben-Or, 1983): the antecedent is a fact of the execution and randomisation buys termination against FLP (Fischer et al., 1985). Here the randomness is in the proposals, so $A(h)$ is a property of the population, not the protocol, and the two uses of randomness are orthogonal. Recent work imports resilient consensus into agent systems (Berdoz et al., 2026; Anand & Pappas, 2026; Ruan et al., 2025) and evaluates resilience against simulated Byzantine agents, prompted (Berdoz et al., 2026; Anand & Pappas, 2026) or replaced by weaker or replayed responses (Lee et al., 2026; Zheng et al., 2026); §6.2 measures how much of an assigned attack a prompted agent carries out, a quantity such evaluations need before they can be compared across models. We use the distinction between logical satisfaction and activation from vacuity analysis (Beer et al., 2001; Kupferman & Vardi, 2003). Strong, differential, median and interval validity (Neiger, 1994; Fitzi & Garay, 2003; Stolz & Wattenhofer, 2015; Melnyk & Wattenhofer, 2018) and the solvability characterisation of Civit et al. (2023) all fix the inputs; our honest-plurality target is a selective (abstaining) form of differential consensus with $\delta = 0$ over *random* inputs, and Theorem 7 makes its attainability depend on the input law.

Agreement in language-model ensembles. Self-consistency aggregates by final answer (Wang et al., 2023); debate and reconciliation let agents revise (Du et al., 2024; Chen et al., 2024); semantic entropy clusters outputs by meaning before measuring uncertainty (Kuhn et al., 2023; Farquhar et al., 2024)—$A(h)$ is the collision (Rényi) form of that entropy on a specified partition, and sampling-consistency checks (Manakul et al., 2023) use collisions at a coarse grain. The decay of $A(h)$ with h reverses the Condorcet jury theorem (Grofman et al., 1983), because validity asks for unanimity rather than a vote; Byzantine-robust aggregation (Blanchard et al., 2017) likewise tolerates faults only when honest inputs are concentrated enough. LLM judges are imperfect instruments whose biases have been documented (Zheng et al., 2023; Wahi, 2026); §6.3 measures which relation they recover and how far it sits from a human reference.

Validity targets and instruments. Xu et al. (2026) give the observed-histogram plurality criterion and observe that the admissibility instrument decides what is certified. We restate the criterion over the compatible set, add the honest-histogram ($2f$) and relabelling forms, and measure how the margin it consumes depends on the grain. Validated agreement uses an external predicate (Cachin et al., 2001); Aegean's Quality Oracle (Ruan et al., 2025) is an instance, and §7 discusses when such an object is needed. Appendix J tabulates the closest prior results, marks which elements each makes explicit, and explains why nearby impossibility results (Liu et al., 2026; Shin, 2026; Scrivens, 2026) do not apply.

3 Problem Formulation

For task x, agent i samples an output string $o_i \sim \pi_i(\cdot \mid x)$. A fixed map g_x assigns it a class $C_i = g_x(o_i)$ in a finite alphabet $\mathcal{C}$; write $p_{i,x}(c) = \Pr[C_i = c \mid x]$ and p when agents are exchangeable. Three objects have different types and the paper keeps them apart: the proposal o_i, a string sampled from $\pi_i(\cdot \mid x)$ by the agent; the class $C_i = g_x(o_i)$, a label in $\mathcal{C}$ fixed by the grain g_x; and the decision $d_i \in \mathcal{C} \cup \{\perp\}$, what a participant outputs under the protocol.

Definition 1 (Grains)**.** The *answer* partition groups normalised final answers. The *answer+KEY* partition groups pairs of the normalised answer and a normalised decisive-step summary (KEY); the prompt requests at most ten words, and on the common task set stored summaries have at most twenty. Answer+KEY is a *lexical* refinement of the answer partition: it records the stated summary string, not the reasoning, and §6.3 measures how far exact string equality is from the human relation "same decisive operation". The KEY is a probe: a short statement of the decisive step that any

automatic grain instrument would have to compare, kept separate from the length and style of the reasoning. We call the two measured partitions the *answer grain* and the *lexical grain*; the human *decisive-operation grain* refines answer classes by annotated operation.

Running example (GSM8K task `gsm_30`, `gpt-4.1-nano`, 30 stored replies)

"Janeth borrowed \$2000 … remaining balance?" All 30 replies answer 220, which is correct. At the answer grain $p = (1)$, so $A(h) = 1$, $\Delta = 1$, $\beta^* = 1/3$, and certification at $n = 10$, $f = 3$ always succeeds. Their KEYs, e.g. *Total owed minus total paid* and *Total owed minus payments made*, split the same replies into 18 lexical classes with modal mass 0.20: $\widehat{A}_U(3) = 0.006$, $\Delta = 0.07$, $\beta^* = 0.03$, and plug-in $f = 3$ coverage $0.2^7 \approx 10^{-5}$. Same replies, same rule: the grain alone moves the guarantee from certain to vacuous.

Among n identified participants at most f are Byzantine, so $h \geq n - f$ are honest; the worst case $h = n - f$ is used throughout. *Agreement*: no two honest participants decide different classes. *Classical validity (CV)*: if every honest participant proposes c, every honest participant decides c. *Weak validity* (a decided value was proposed by some honest participant) has no unanimity antecedent and hence no vacuity, but for agents with non-zero individual error it licenses deciding one honest participant's mistake. *Honest-plurality validity* (soundness form): an honest participant that decides $c \neq \perp$ decides the strict-plurality class of the honest proposals; abstention is allowed. Whenever it decides, it satisfies CV, since a unanimous honest class is the strict plurality; it is the target a certificate can check from observed labels, and the thresholds below concern it, not CV, which classical protocols satisfy whatever $A(h)$. *External validity*: every honest decision is accepted by an external predicate $V(x, \cdot)$. *Termination*: every honest participant decides. The antecedent of CV is

$$\mathcal{U}_h = \{\exists c : \; C_1 = \cdots = C_h = c\}, \tag{1}$$

a proposal-level event that says nothing directly about decisions.

Assumption 1. (i) Given the task and any context Z shared by the honest agents, honest labels are i.i.d. from p_Z; except in Proposition 8, Z is trivial. (ii) Byzantine participants may choose labels after seeing the honest ones and act by appending at most f labels. (iii) The decision layer observes a common histogram $N = (N_c)_c$, produced by a reliable-broadcast layer that we do not supply; inconsistent views are out of scope (Appendix H).

4 The Validity Gap

4.1 Activation: unanimity is a collision event

Proposition 1 (Collision identity). *For $h \geq 1$ conditionally independent, identically distributed honest labels,*

$$\Pr[\mathcal{U}_h] = A(h) := \textstyle\sum_c p(c)^h \leq p_{\max}^{h-1}, \qquad p_{\max} = \max_c p(c), \tag{2}$$

with equality for $h > 1$ iff p is uniform on its support. If p is non-degenerate, $A(h) \to 0$ geometrically. Equivalently $A(h) = \exp\{-(h-1)H_h(p)\}$ with H_h the Rényi entropy of order h (Rényi, 1961).

Independence does real work: if all h agents copy one shared draw, each marginal is p yet $A(h) = 1$.

Proposition 2 (Vacuity is not infeasibility). *(i) A rule that ignores all proposals and always decides d satisfies CV on a sampled execution with probability $1 - A(h) + p(d)^h \geq 1 - A(h)$. (ii) If $f = 0$ and every message is eventually delivered, deciding the proposal of the smallest-identifier participant satisfies Agreement, CV and Termination on every execution, while $A(n)$ may be arbitrarily small.*

Part (i) says a rare antecedent makes the clause *easy* to satisfy; part (ii) says agreement is unaffected, and for $f > 0$ any classical Byzantine agreement protocol with $n > 3f$ likewise satisfies CV on every execution, whatever $A(h)$. What changes is force, not feasibility; adding agents does not help, since at $f = 0$ the relevant quantity is $A(n)$, which falls with n.

Proposition 3 (Refinement and mixtures). *If $\mathcal{C}'$ refines $\mathcal{C}$ then $A_{\mathcal{C}'}(h) \leq A_{\mathcal{C}}(h)$ for every h. The top-two margin $p_{\max} - p_2$ is not monotone under refinement: $(0.5, 0.3, 0.2)$ has margin 0.2; splitting the mode gives $(0.3, 0.2, 0.3, 0.2)$ with margin 0, splitting the runner-up into $0.15 + 0.15$ gives margin 0.3. For independent agents with laws $p_1, \ldots, p_h$, $A(h) = \sum_c \prod_i p_i(c) \leq \sum_c \bar{p}(c)^h$ with $\bar{p} = \frac{1}{h} \sum_i p_i$.*

4.2 CERTIFICATION: FROM HONEST AGREEMENT TO OBSERVABLE SUPPORT

The clause quantifies over honest participants; an observer sees everyone. A shared profile of nine votes for a and one for b with $f = 1$ is compatible with two executions: if the b-proposer is faulty, all honest proposals are a and CV requires a; if an a-proposer is faulty, the honest proposals are mixed and the antecedent is false. All-participant unanimity is therefore not a test of honest unanimity, and a rule must reason over compatible executions. Let $U_f(N)$ be the histograms obtained by deleting at most f entries from N, $T(H)$ the strict-plurality class of H (empty on ties), and $K_f(N) = \bigcap_{H \in U_f(N)} T(H)$.

Theorem 4 (Exact margin certificate; proofs in Appendix A)**.** *For $0 \le f < n$, $\mathrm{certify}(N, f) = c \iff N_c > f + \max_{d \neq c} N_d$ returns the unique element of $K_f(N)$ and otherwise abstains (the criterion is that of Xu et al., 2026, Thm. 5.7). If up to e remaining entries may be arbitrarily relabelled, the guard becomes $N_c > f + 2e + \max_{d \neq c} N_d$.*

Beyond this criterion we add the relabelling guard, the honest-histogram requirement (necessary for every sound rule), the threshold of Theorem 7 and their grain dependence.

Proposition 5 (Honest margin)**.** *Let H be the honest histogram and $N = H + B$ with $|B| \le f$ appended labels. Then c^* is certified for every allowed B iff*

$$H_{c^*} - \max_{d \neq c^*} H_d > 2f, \tag{3}$$

one f consumed by the attack's reduction of the lead and one required by the certificate. The condition is necessary for every rule that is sound for honest-plurality validity under a common view: such a rule can only decide an element of $K_f(N)$, which the rule of Theorem 4 decides whenever it exists.

Remark 6 (Threshold degeneracy). Counts are integers, so Eq. (3) reads $H_{c^*} - \max_{d \neq c^*} H_d \ge 2f+1$. Any honest dissent makes $\max_{d \neq c^*} H_d \ge 1$ and hence needs $H_{c^*} \ge 2f+2$ and $h \ge 2f+3$, i.e. $n \ge 3f + 3$. For $3f + 1 \le n \le 3f + 2$, which includes the classical threshold and the setting $n = 10$, $f = 3$ used below ($h = 7$), the certificate requires every honest label on c^*, and its coverage (the probability that c^* is certified) is exactly $\Pr[C_1 = \cdots = C_h = c^*] = p(c^*)^h$: at the threshold the certification gap is the activation gap on the mode. For $n \le 3f$ ($h \le 2f$), even honest unanimity cannot meet Eq. (3), so coverage is zero.

Theorem 7 (Dispersion-limited resilience)**.** *Let honest labels be i.i.d. from p with unique mode c^* and top-two margin $\Delta = p(c^*) - \max_{d \neq c^*} p(d)$, let $f = \lfloor \beta n \rfloor$ for a fault fraction $0 \le \beta < 1$, $h = n - f$, and $\beta^* = \Delta/(2 + \Delta)$. Let P_n be the largest probability, over rules sound for honest-plurality validity under a common view, that c^* is certified against every append attack. With $m - 1$ the support size of p,*

$$\begin{aligned} &\beta < \beta^* : && P_n \ge 1 - (m-1)\, e^{-\frac{h}{2}(\Delta - 2f/h)^2} \to 1, \\ &\beta > \beta^*,\ 2f \ge h\Delta : && P_n \le e^{-\frac{h}{2}(2f/h - \Delta)^2} \to 0, \end{aligned} \tag{4}$$

where $2f \ge h\Delta$ holds for all large n when $\beta > \beta^*$. Hence β^* is a sharp threshold: certification probability tends to one below it and to zero above it, and $\beta^* \le 1/3$ with equality iff $\Delta = 1$.

The classical requirement $n > 3f$ is the special case of labels that are deterministic at the grain ($\Delta = 1$). For stochastic agents the tolerable fault fraction is set by the honest margin at the chosen grain, and no number of agents raises it: adding agents drives coverage to one below β^* and to zero above it (Fig. 4, Appendix B). Refinement can move Δ either way (Proposition 3), so β^* has to be reported per grain. Equivalently, certification needs $\Delta > 2\beta/(1 - \beta)$: $\Delta > 0.5$ at $\beta = 0.2$ and $\Delta > 0.22$ at $\beta = 0.1$; the boundary $\beta = \beta^*$ is not covered.

Proposition 8 (Shared context)**.** *Under Assumption 1 with a random shared context Z and marginal law $p = \mathbb{E}_Z p_Z$, $A(h) = \mathbb{E}_Z \sum_c p_Z(c)^h \ge \sum_c p(c)^h$. If each p_Z has a unique mode and $\Pr_Z[\beta^*(\Delta_Z) = \beta] = 0$, the largest probability that the mode of p_Z is certified against every append attack tends to $\Pr_Z[\beta^*(\Delta_Z) > \beta]$.*

Shared prompts or retrieved context thus raise activation but make the threshold random; our round-0 replies are separate calls, consistent with a trivial Z (Table 4, Appendix F). Abstention is a completed selective decision, not a termination failure.

5 MEASUREMENT SETUP

We use stored replies from `gpt-4.1-nano`, `gpt-4o-mini` and `gpt-4.1-mini` at temperature 1.0, 30 requested per model and task; the common set is 123 tasks (50 GSM8K, 26 MMLU, 47 MBPP; Cobbe et al., 2021; Hendrycks et al., 2021; Austin et al., 2021) on which every model has at least ten parseable replies. Scoring is benchmark-specific (numeric for GSM8K, option letter for MMLU, typed Python literal for MBPP); KEY normalisation is cosmetic only; an unparseable answer is missing, never a class (Appendix C). With k replies and class counts n_c, $A(h)$ is estimated by the U-statistic $\widehat{A}_U(h) = \sum_c [n_c]_h / [k]_h$, $[a]_h = a(a-1)\cdots(a-h+1)$, unbiased under i.i.d. sampling and undefined for $k < h$; the plug-in $\sum_c (n_c/k)^h$ over-states $A(10)$ at the lexical grain by a factor 5.2 (`gpt-4.1-nano`) (Appendix D). Every headline mean gives each task equal weight; intervals are percentile bootstraps over tasks, paired where conditions share tasks.

6 EXPERIMENTS: THREE VIEWS OF THE GAP

6.1 Q1: HOW DOES THE GRAIN CHANGE ACTIVATION?

Prediction (Propositions 1 and 3): $A(h)$ falls geometrically at both grains, and answer+KEY lies below answer at every h.

Final answers are concentrated (Table 1; Fig. 5 in Appendix D): mean modal mass is 0.928–0.977 across models. Keeping the decisive-step string lowers it to 0.147–0.164, with 22–24 observed classes per task; the same ordering holds on four further models from three vendors (Appendix E). For `gpt-4.1-nano`, moving from the answer grain to the lexical grain lowers $\widehat{A}_U(3)$ from 0.855 to 0.0125 and $\widehat{A}_U(10)$ from 0.753 to 2.7×10^{-5}. Humans rate far more pairs equivalent (§6.3), so the ordering of Proposition 3 carries over to the human relation but the magnitude does not.

The prediction holds in live groups with nothing fitted. Benign debate (Du et al., 2024) on `gpt-4.1-nano`, 50 GSM8K tasks, $n \in \{3, 5, 7, 10\}$, five repetitions. Round-0 replies are sampled before any message, so the collision identity applies; later rounds are communication-coupled and are reported separately. Single-agent histograms on the same tasks predict the observed round-0 unanimity within one percentage point at every n (Fig. 2a; Table 4, Appendix F). Agreement falls with n from 0.87 to 0.76—the trend reported by Berdoz et al. (2026); Anand & Pappas (2026) likewise find that benign groups often fail to reach consensus. Round-0 replies are independent calls, so this checks the collision identity; it does not explain their post-interaction results.

The product law predicts mixed groups too. Proposition 3 extends the prediction to a mixed group of three `gpt-4.1-nano`, three `gpt-4o-mini` and four `gpt-4.1-mini` agents: predicted 0.767, observed 0.777 over 247 debates. The prediction lies between the homogeneous predictions for nano (0.748) and 4.1-mini (0.938).

Interaction moves agreement more than accuracy, but the data do not separate the two. At $n = 10$, two rounds raise answer unanimity from 0.756 to 0.832 and majority accuracy (the plurality answer is correct) from 0.916 to 0.932. The paired difference of the two gains, 6.0 points ($[-1.2, 12.8]$), is not separated from zero; only the $n = 10$ agreement gain excludes zero (Fig. 2b; Appendix F). Round-2 lexical unanimity is 0.004: answers align, stated summaries do not.

Table 1: Dispersion on the common 123-task set; per-task statistics averaged over tasks; all $\widehat{A}$ are U-statistics with $k \geq h$.

model	partition	$p_{\max}$	p_2	classes	accuracy	$\widehat{A}_U(3)$	$\widehat{A}_U(10)$
`gpt-4.1-nano`	answer	0.928	0.047	1.7	0.872	0.855	0.753
	answer+KEY	0.147	0.080	23.6	0.872	0.0125	2.7×10^{-5}
`gpt-4o-mini`	answer	0.931	0.054	1.5	0.878	0.862	0.778
	answer+KEY	0.159	0.090	21.9	0.878	0.0133	2.4×10^{-5}
`gpt-4.1-mini`	answer	0.977	0.021	1.1	0.937	0.951	0.915
	answer+KEY	0.164	0.084	23.0	0.937	0.0256	0.00578

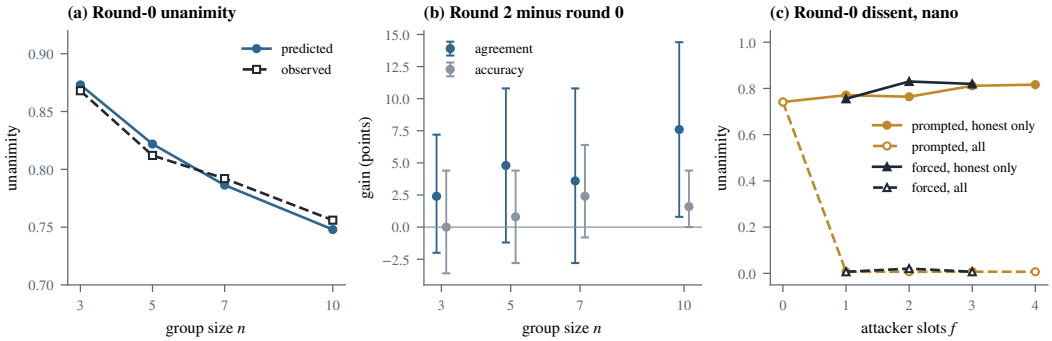

Figure 2: **Live groups: prediction (Q1), interaction and dissent (Q2).** (a) Predicted vs. observed round-0 unanimity. (b) Paired round-2 minus round-0 gains with 95% task-bootstrap intervals. (c) Round-0 unanimity at $n = 10$ on `gpt-4.1-nano`, same 50 tasks, about 145 trials per f: honest-only and all-participant, with a prompted or a forced attacker. Round-2 counterparts are in Table 5 (Appendix F).

Q1 findings: activation

Answers are concentrated (modal mass 0.93–0.98) but stated decisive steps are not (0.15–0.16, 22–24 classes per task): on `gpt-4.1-nano` $\widehat{A}_U(10)$ falls from 0.753 to 2.7×10^{-5}. Single-agent samples predict round-0 group unanimity within one point for $n = 3$–10, and a mixed 3/3/4 group (0.767 vs 0.777), with no fitted parameter.

6.2 Q2: What can an observer certify?

Prediction (Theorem 4, Proposition 5, Theorem 7): the certifiable budget is set by the honest top-two margin. Since the mean margin falls from about 0.9 at the answer grain to below 0.1 at the lexical grain (Table 1), the classical budget should survive with high probability at the former and rarely at the latter (lexical values are instrument-relative; §6.3).

One dissenter hides the antecedent; how much it dissents depends on the model. With one `gpt-4.1-nano` slot assigned a fixed wrong answer, round-0 all-participant unanimity falls to 0.007 under either attacker defined next, while honest-only unanimity is 0.771 (prompted) and 0.755 (forced), close to the predicted $A(9) \approx 0.76$ (Fig. 2c). A *forced* attacker has its wrong value written into its messages; a *prompted* one is only instructed to argue for it. Both run at $n = 10$, $f \in \{1, 2, 3\}$, three rounds, on `claude-opus-5` (24 tasks, 48 trials per cell) and `gpt-4.1-nano` (tasks of Table 4, 144–146 trials; Table 5, Appendix F). In round 2, forced attackers hold all-participant unanimity at ≤ 0.021, while honest unanimity stays at 0.958–1.000 (opus) and 0.855–0.875 (nano). Prompted attackers leave it at 0.438–0.667 (opus) and 0.083–0.200 (nano): the opus attacker adopts its target in $10/288$ slots (3.5%), the nano attacker in 74–78%. This is a two-model pilot with collection times that are not aligned (Appendix F), not a cross-model law. Adoption reflects instruction following, not the protocol, so it must be reported beside any tolerance result, as would help prior evaluations with prompted attackers (Berdoz et al., 2026; Anand & Pappas, 2026); the forced attacker is a model-independent reference.

Dispersion sets the certifiable budget. Each task's measured law is taken as the honest law; 30,000 honest histograms per setting are drawn and Eq. (3) evaluated on the full alphabet (Appendix B). At the classical budget $f = 3$, the designated (empirical) mode is certified with probability 0.843 at the answer grain versus 0.004 at the lexical grain (plug-in; unbiased 0.836 and 0.003); at $f = 0$, the corresponding probabilities are 0.970 and 0.225 (Fig. 3a, Appendix B). By Remark 6, the $f = 3$ event is honest unanimity on the mode, $\mathbb{E}_{\text{task}}[p(c^*)^7]$, so even at the answer grain, whose labels humans confirm on 99.4% of pairs, it fails on 16% of draws. At the lexical grain the rule abstains in 42% of draws even with no faults (Table 2).

The tolerable fault fraction per grain. Theorem 7 turns each measured law into $\beta^* = \Delta/(2+\Delta)$. At the answer grain the mean over model–task cells is 0.306, and 76% of cells reach $1/3$ because all replies agree (a plug-in, hence optimistic, value). At the lexical grain the mean is 0.032 and no

cell reaches $1/3$. At $\beta = 0.2$ and $n = 160$, the certified share is 0.912 at the answer grain and 0.021 lexically, matching the shares of cells with $\beta^* > 0.2$ (0.916 and 0.022; Fig. 4, Appendix B).

Q2 findings: certification

At $n = 10$, $f = 3$ certification is honest unanimity on the mode: coverage 0.843 at the answer grain, 0.004 lexically. The tolerable fault fraction β^* averages 0.31 at the answer grain and 0.03 lexically. Prompted attackers adopt their target in 3.5% (opus) to 74–78% (nano) of round-2 slots.

6.3 Q3: What relation do agreement instruments measure?

Prediction (Definition 1, Remark 10 in Appendix A): if instruments do not recover the intended relation, the measured margin is instrument-relative, candidate instruments should disagree, and their errors should have a direction.

Instruments disagree. We take 800 within-task `gpt-4.1-nano` pairs from 147 tasks of the full 150-task pool (722 with the same final answer, 78 with different answers) and ask four instruments whether two replies, each given as final answer and KEY, share the same decisive step: a lexical predicate, an embedding threshold (`text-embedding-3-small`, cosine 0.9) and two judges (`gpt-4.1-nano`, `gpt-4.1-mini`). They return SAME on 0.035 $[0.020, 0.053]$, 0.220 $[0.177, 0.263]$, 0.909 $[0.874, 0.940]$ and 0.890 $[0.854, 0.924]$ of pairs (Fig. 6a, Appendix G). The judges follow the final answer: the nano judge says SAME on 0.989 of same-answer pairs and 0.167 $[0.075, 0.275]$ of different-answer pairs, the mini judge on 0.982 and 0.038. The lexical rule says SAME on 0.039 and 0.000 (Fig. 6b).

Human calibration. A human reference covering all models and pair types uses a separate frame. Three annotators labelled 300 within-task pairs, a pre-specified stratified sample of all $518{,}837$ stored pairs with inverse-probability weights, for same answer and same decisive operation (SAME/DIFF/VAGUE/UNSURE; wording, language and symbolic form never count as a difference; Appendix G). Like the instruments, annotators see final answers and KEYs, not full reasoning, so the human relation is judged on the stated decisive step. Fleiss' κ is 0.947 on answers and 0.532 on operations. The answer normaliser errs on 0.6% of weighted pairs. The lexical rule has weighted error 0.751 $[0.693, 0.807]$, false rejection 0.967 $[0.963, 0.970]$ and no observed false acceptance.

Human-calibrated pairwise agreement. On same-model calibration pairs, weighted pairwise agreement is 0.952 under human answer labels (0.948 under answer equality), 0.043 under lexical equality and 0.67–0.85 under the human answer+operation reference, counting unresolved pairs as DIFF or SAME (Table 10, Appendix G); the range is not a confidence interval. Relative to answer agreement, the human reference lowers pairwise agreement by a factor of 1.1–1.4, not the 22-fold of the lexical rule: lexical over-splitting accounts for most of the collapse. Pairwise rates do not define a partition or its modal mass.

Certification at the decisive-operation grain. If the human relation is a partition refining the answer partition, pairwise human rates bound coverage at $n = 10$, $f = 3$ to $[0.06, 0.79]$ (Table 8, Appendix G). Direct partition annotation of 24 model–task cells then places coverage at 0.43 with an unbiased estimator (95% interval over cells and annotators $[0.27, 0.61]$), inside these bounds, against 0.82 at the answer grain and 0.0001 lexically on the same replies; the paired answer–operation gap is 0.39 $[0.22, 0.55]$ (Table 9, Appendix G).

Instruments against the human reference. Against the resolved human reference (Table 7, Appendix G), both judges rarely reject falsely (≤ 0.01) but falsely accept 0.583–0.666 of human-DIFF pairs, as answer equality alone does (0.583); their weighted errors (0.138–0.149) do not beat its 0.132. The judges agree with answer equality on 194 (nano) and 202 (mini) of the 204 resolved pairs. Prompting does not repair this: a few-shot variant is unchanged (false acceptance 0.685) and a chain-of-thought variant raises it to 0.842. The embedding threshold errs on the strict side (false rejection 0.722, false acceptance 0.009). The judges thus compute verdict equality and the lexical rule computes string equality; neither recovers the human operation relation between them.

Q3 findings: instruments

Four instruments return SAME on 0.035–0.909 of the same pairs. The lexical rule rejects 97% (weighted) of human-equivalent pairs; judges agree with answer equality on 194–202 of 204 pairs. On 24 annotated cells, human operation partitions give coverage 0.43 $[0.27, 0.61]$ against 0.82 at the answer grain.

7 DISCUSSION

An external checking target. An executable checker supplies validity from outside the population and sidesteps the instrument gap: in an MBPP case study with `claude-opus-5` (Appendix F) the group finds an accepted candidate in 0.90–0.92 of trials and rejects every constructed wrong return, at no cost in fault budget. Aegean's Quality Oracle (Ruan et al., 2025) is one such target. The operation grain matters when the output is a method rather than a value, such as an audited derivation or a reviewed code patch; the answer grain suffices when only the value is consumed.

Diversity is paid for in activation. Heterogeneous groups are recommended to decorrelate faults (Bhardwaj et al., 2026); the collision identity prices this. A balanced mixture of the three models lowers answer-grain $\widehat{A}_U(9)$ from 0.918 for the best single model to 0.735, so the validity clause applies less often. Mixing may still help correctness: the live mixed group of §6.1 is more accurate at round 0 than homogeneous nano (0.931 against 0.916), which our data cannot separate from the stronger members it contains.

Design implications and open problems. A semantic layer could emit candidates with external certificates and a protocol layer run validated agreement (Cachin et al., 2001) over certified digests; since distinct valid outputs have distinct digests, that layer must accept any certified digest rather than wait for unanimity, and substituting a judge for the certificate re-imports the instrument gap. Five problems remain open: validity conditions that price the shared context Z of Proposition 8 rather than assume it away; measuring decisive-operation grains at scale, for example by fusing instruments of known error direction around a small human anchor; adversaries that choose the grain at which they deviate; treating the requested output form as a design parameter that coarsens g_x without merging decisions that matter; and the same accounting for strong, median and interval validity, which also quantify over honest proposals.

Reporting checklist for consensus claims over stochastic agents

(1) the grain g_x and the instrument that implements it; (2) $A(h)$ at that grain, estimated by a U-statistic; (3) the label source, its calibration against the intended relation, and the direction of its error; (4) what the adversary actually emitted, not what it was told; (5) which of dispersion and fault count limits the certificate; (6) accuracy reported beside agreement, never as a proxy. Appendix I gives an instance of each item from this paper and the error its omission causes.

8 LIMITATIONS

The answer+KEY partition is lexical. Calibration shows the lexical rule splits 97% (weighted) of resolved pairs judged equivalent on both the answer and decisive-operation axes, so the lexical activation and certification numbers measure dispersion under a stated instrument, not semantic dispersion. The defensible fine-grained claims are the ordering of Proposition 3, the instrument disagreement of §6.3, the human pairwise range $[0.67, 0.85]$ and the partition-based coverage estimates. The decisive-operation estimates are judged from stated KEYs, cover 24 annotated model–task cells of 12 replies each and assume that the human relation is a partition; humans agree on it only moderately ($\kappa = 0.532$, ARI 0.52), and no automatic instrument we tested recovers it. The main experiments use three small models from one vendor; the ordering of grains replicates on four further models from three vendors (Appendix E). Appendix H covers further limits.

9 CONCLUSION

For stochastic agents, classical validity has a random antecedent whose probability depends on the agreement grain. Certifying the honest mode needs an honest count margin above $2f$, so the tolerable fault fraction is $\Delta/(2+\Delta)$, below one third unless labels are deterministic. At $n = 10$, $f = 3$ even the answer grain certifies with mean probability only 0.843, and on 24 annotated cells the decisive-operation grain roughly halves this. Exact strings over-split and judges track the answer, so guarantees should name the grain, report $A(h)$ and calibrate the instrument. Agreement, validity and correctness are distinct quantities, and a consensus claim over stochastic agents is informative only when each is reported at a stated grain.

Reproducibility statement

All reported numbers are computed offline from archived model replies, debate transcripts and annotation sheets provided as supplementary material (Appendix K). One script reruns every analysis without API access and checks the reported numbers against the result files; the formal results of §4 are proved in Appendix A and checked by exhaustive enumeration on small instances. Prompts, model identifiers and sampling settings are stored with the replies; collecting new replies requires API access.

AI use statement

Generative AI tools were used for language polishing and pre-submission feedback on the theoretical scope, experimental methodology, interpretation of results, and manuscript presentation. They also assisted with wording revisions. The authors take responsibility for all claims, analyses, and the final manuscript.

References

Sribalaji C. Anand and George J. Pappas. Resilient consensus in agentic AI. *arXiv preprint arXiv:2606.15024*, 2026.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Ilan Beer, Shoham Ben-David, Cindy Eisner, and Yoav Rodeh. Efficient detection of vacuity in temporal model checking. *Formal Methods in System Design*, 18(2):141–163, 2001. doi: 10.1023/A:1008779610539.

Michael Ben-Or. Another advantage of free choice: Completely asynchronous agreement protocols. In *PODC*, pp. 27–30, 1983.

Frédéric Berdoz, Leonardo Rugli, and Roger Wattenhofer. Can AI agents agree? *arXiv preprint arXiv:2603.01213*, 2026.

Varun Pratap Bhardwaj, Garima Singh, and Arun Pratap Bhardwaj. Agent behavioral contracts II: Certifying compositional reliability without assuming independence. *arXiv preprint arXiv:2608.12895*, 2026.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, pp. 119–129, 2017.

Christian Cachin, Klaus Kursawe, Frank Petzold, and Victor Shoup. Secure and efficient asynchronous broadcast protocols. In *Advances in Cryptology – CRYPTO 2001*, volume 2139 of *LNCS*, pp. 524–541. Springer, 2001.

Miguel Castro and Barbara Liskov. Practical Byzantine fault tolerance. In *OSDI*, pp. 173–186, 1999.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *ACL*, pp. 7066–7085, 2024.

Pierre Civit, Seth Gilbert, Rachid Guerraoui, Jovan Komatovic, and Manuel Vidigueira. On the validity of consensus. In *Proceedings of the 42nd ACM Symposium on Principles of Distributed Computing (PODC)*, pp. 332–343, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Danny Dolev and H. Raymond Strong. Authenticated algorithms for Byzantine agreement. *SIAM Journal on Computing*, 12(4):656–666, 1983.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *ICML*, volume 235 of *PMLR*, pp. 11733–11763, 2024.

Cynthia Dwork, Nancy Lynch, and Larry Stockmeyer. Consensus in the presence of partial synchrony. *Journal of the ACM*, 35(2):288–323, 1988.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Michael J. Fischer, Nancy A. Lynch, and Michael S. Paterson. Impossibility of distributed consensus with one faulty process. *Journal of the ACM*, 32(2):374–382, 1985.

Matthias Fitzi and Juan A. Garay. Efficient player-optimal protocols for strong and differential consensus. In *Proceedings of the 22nd ACM Symposium on Principles of Distributed Computing (PODC)*, pp. 211–220, 2003.

Bernard Grofman, Guillermo Owen, and Scott L. Feld. Thirteen theorems in search of the truth. *Theory and Decision*, 15(3):261–278, 1983.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*, 2023.

Orna Kupferman and Moshe Y. Vardi. Vacuity detection in temporal model checking. *International Journal on Software Tools for Technology Transfer*, 4(2):224–233, 2003.

Leslie Lamport, Robert Shostak, and Marshall Pease. The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.

Haejoon Lee, Vincent-Daniel Yun, Dimitra Panagou, and Sai Praneeth Karimireddy. Robust multi-agent LLMs under Byzantine faults. *arXiv preprint arXiv:2605.09076*, 2026.

Christian List and Philip Pettit. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, 18(1):89–110, 2002.

Jiayuan Liu, Shiyi Du, Weihua Du, Mingyu Guo, and Vincent Conitzer. The consensus trap: Rescuing multi-agent LLMs from adversarial majorities via token-level collaboration. *arXiv preprint arXiv:2604.17139*, 2026.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023.

Darya Melnyk and Roger Wattenhofer. Byzantine agreement with interval validity. In *Proceedings of the 37th IEEE Symposium on Reliable Distributed Systems (SRDS)*, pp. 251–260, 2018.

Gil Neiger. Distributed consensus revisited. *Information Processing Letters*, 49(4):195–201, 1994.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 547–561. University of California Press, 1961. URL `https://static.renyi.hu/renyi_cikkek/1961_on_measures_of_entropy_and_information.pdf`.

Chaoyi Ruan, Yiliang Wang, Ziji Shi, and Jialin Li. Reaching agreement among reasoning LLM agents. *arXiv preprint arXiv:2512.20184*, 2025.

Arsenios Scrivens. Information-theoretic limits of safety verification for self-improving systems. *arXiv preprint arXiv:2603.28650*, 2026.

Kwan Soo Shin. The reasoning trap: An information-theoretic bound on closed-system multi-step LLM reasoning. *arXiv preprint arXiv:2605.01704*, 2026.

David Stolz and Roger Wattenhofer. Byzantine agreement with median validity. In *Proceedings of the 19th International Conference on Principles of Distributed Systems (OPODIS)*, volume 46 of *LIPIcs*, pp. 22:1–22:14, 2015.

Vansh Wahi. LLM-as-a-judge is not an oracle: Why self-improving agents need deterministic guardrails. *arXiv preprint arXiv:2609.02246*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

Haoran Xu, Lei Zhang, Iadh Ounis, and Xianbin Wang. Certifiable semantic agreement among LLM agents: What the admissibility instrument decides. *arXiv preprint arXiv:2606.07316*, 2026. URL https://arxiv.org/abs/2606.07316v2.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS Datasets and Benchmarks*, 2023.

Lifan Zheng, Jiawei Chen, Qinghong Yin, Jingyuan Zhang, Xinyi Zeng, and Yu Tian. Rethinking the reliability of multi-agent system: A perspective from Byzantine fault tolerance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35012–35020, 2026. doi: 10.1609/aaai.v40i41.40806.

Appendix overview. Appendix A proves the results of §4. Appendix B details the certification computation of §6.2, Appendix C the answer scoring and Appendix D the dispersion curves and estimators of §6.1, and Appendix E their replication on further vendors. Appendix F reports the live-group, dissent and checker runs, and Appendix G the instrument comparison, human calibration and decisive-operation partitions of §6.3. Appendices H–K give further limitations, the reporting checklist, the closest prior work and code availability.

A PROOFS

Proposition 1. Conditional independence gives $\Pr[C_1 = \cdots = C_h = c] = p(c)^h$; the events are disjoint over c. Since $p(c)^h \le p(c)p_{\max}^{h-1}$, summing gives the bound; for $h > 1$ equality needs every positive $p(c)$ to equal $p_{\max}$. Non-degeneracy gives $p_{\max} < 1$.

Proposition 2. (i) The fixed rule violates CV exactly when the honest proposals are unanimous in some $c \neq d$, probability $A(h) - p(d)^h$. (ii) With $f = 0$ every participant eventually receives the designated proposal and decides it (Agreement, Termination); if all propose c the designated one does too (CV). Uniform p on m classes gives $A(n) = m^{1-n}$.

Proposition 3. $(a + b)^h \ge a^h + b^h$ for $a, b \ge 0$, applied within each coarse class. The margin examples are checked directly. AM–GM gives $\prod_i p_i(c) \le \bar{p}(c)^h$.

Theorem 4. ($\Leftarrow$) If $N_c > f + \max_{d \neq c} N_d$, every compatible H has $H_c \ge N_c - f > N_d \ge H_d$. ($\Rightarrow$) If the guard fails, delete $\min(f, N_c)$ copies of c: $H_c = \max(0, N_c - f) \le N_{d^*} = H_{d^*}$ for a maximal rival d^*. Relabelling: the compatible set is $\mathcal{U}_{f,e}(N) = \{H : \exists N' \le N,\ |N - N'|_1 \le f,\ |H|_1 = |N'|_1,\ \frac{1}{2}|N' - H|_1 \le e\}$ and $\mathrm{certify}(N, f, e) = c \iff \bigcap_{H \in \mathcal{U}_{f,e}(N)} T(H) = \{c\}$. Deleting reduces a pairwise lead by at most one per entry, moving an entry from c to a rival by two; e arbitrary relabellings cost at most $2e$, attained by moving copies of c to the maximal rival.

Proposition 5. An append budget f reduces the lead of c^* by at most f, attained by appending to its strongest rival; the observed guard then needs a further f, giving Eq. (3). Conversely, if Eq. (3) fails, appending f copies of the strongest rival makes the guard fail. For any rule R sound for honest-plurality validity, if $R(N) = c \neq \perp$ then $c = T(H)$ for every honest histogram H compatible with N, so $c \in K_f(N)$; by Theorem 4, $\mathrm{certify}(N, f) = c$ as well. Hence no sound rule certifies c^* on an input where certify abstains, and the necessity of Eq. (3) transfers to every sound rule.

Coverage bound used by Theorem 7.

Lemma 9 (Coverage is a dispersion statement)**.** *Let honest labels be i.i.d. from p with unique mode c^*, m one more than the support size of p (all unobserved classes have honest count zero, so together they act as one rival), and $\Delta = p(c^*) - \max_{d \neq c^*} p(d)$. If $h\Delta > 2f$,*

$$\Pr[c^* \textit{ certified for every allowed } B] \ \ge\ \max\Big\{0,\ 1 - (m-1)\exp\Big(-\tfrac{(h\Delta - 2f)^2}{2h}\Big)\Big\}. \qquad (5)$$

Lemma 9. For rival d let $Z_{i,d} = \mathbb{1}[C_i = c^*] - \mathbb{1}[C_i = d] \in [-1, 1]$, $D_d = \sum_i Z_{i,d}$, $\delta_d = p(c^*) - p(d) \ge \Delta$. Hoeffding (1963) gives $\Pr[D_d \le 2f] \le \exp\{-(h\delta_d - 2f)^2/(2h)\}$; a union bound over the $m - 1$ rivals and $\delta_d \ge \Delta$ give Eq. (5).

Theorem 7. By Proposition 5 and its extension to sound rules, $P_n = \Pr[H_{c^*} - \max_{d \neq c^*} H_d > 2f]$, attained by the rule of Theorem 4. If $\beta < \beta^*$ then $\Delta > 2\beta/(1 - \beta) \ge 2f/h$, and Lemma 9 gives the lower bound; its exponent grows linearly in n. If $\beta > \beta^*$, let d_2 be the runner-up class; certification requires $D = H_{c^*} - H_{d_2} > 2f$, where D is a sum of h independent terms in $[-1, 1]$ with mean $h\Delta$. Since $f/h \to \beta/(1 - \beta) > \Delta/2$, eventually $2f > h\Delta$ and Hoeffding's inequality gives $\Pr[D > 2f] \le \exp\{-(2f - h\Delta)^2/(2h)\}$. Finally $\beta < \beta^* \iff (1 - \beta)\Delta > 2\beta$, and $\Delta/(2 + \Delta)$ is increasing on $[0, 1]$ with value $1/3$ at $\Delta = 1$.

Proposition 8. Conditional independence given Z gives $A(h) = \mathbb{E}_Z \sum_c p_Z(c)^h$, and convexity of $x \mapsto x^h$ with Jensen's inequality gives $\mathbb{E}_Z p_Z(c)^h \ge p(c)^h$. Soundness is a property of each

execution, so the rule of Theorem 4 remains optimal given Z, and the certification probability is $\mathbb{E}_Z[P_n(p_Z)]$. By Theorem 7, $P_n(p_Z) \to \mathbb{1}[\beta^*(\Delta_Z) > \beta]$ whenever $\beta^*(\Delta_Z) \neq \beta$, and dominated convergence gives the limit.

Remark 10 (Instrument error and the margin). The relabelling form of Theorem 4 quantifies how label errors affect certification: on a fixed alphabet, relabelling at most e entries changes a candidate's lead over its largest rival by at most $2e$, giving the guard $N_c > f + 2e + \max_{d \neq c} N_d$. Moving entries from the candidate to a rival erodes its support; merging rival entries into it inflates its support. The calibration in §6.3 diagnoses lexical over-splitting and judge over-acceptance, but its pairwise error rates are not proposal-level relabelling rates and cannot be substituted for e. The measured histograms, not this conversion, establish the certification gap.

B Certification computation

For each model, task and grain, normalised stored counts define a multinomial law and the empirical mode is designated (ties broken deterministically; the event has the same probability for every tied class); 30,000 honest histograms per setting are drawn with a recorded seed and Eq. (3) evaluated, with one zero-probability rival appended so the adversary may introduce an unobserved label. The aggregate is the mean of the model-specific probabilities within each task, averaged equally over tasks (Table 2). These are plug-in coverages; at $f = 3$ the unbiased estimates $\mathbb{E}_{\text{task}}[[n_{c^*}]_7/[k]_7]$ are 0.836 (answer) and 0.003 (answer+KEY). Certifying *some* class at $f = 0$ has probability 0.987 (answer) and 0.582 (answer+KEY), since a different class can win a realised sample (Fig. 3b).

Table 2: Certification probabilities at $n = 10$, $h = 10 - f$; means over tasks with 95% task-bootstrap intervals.

event	answer	answer+KEY
designated mode, $f = 0$	0.970 [0.956, 0.983]	0.225 [0.198, 0.255]
designated mode, $f = 1$	0.948 [0.925, 0.968]	0.069 [0.053, 0.087]
designated mode, $f = 2$	0.897 [0.863, 0.929]	0.018 [0.011, 0.027]
designated mode, $f = 3$	0.843 [0.799, 0.885]	0.004 [0.001, 0.009]
any class, $f = 0$	0.987 [0.981, 0.992]	0.582 [0.568, 0.597]

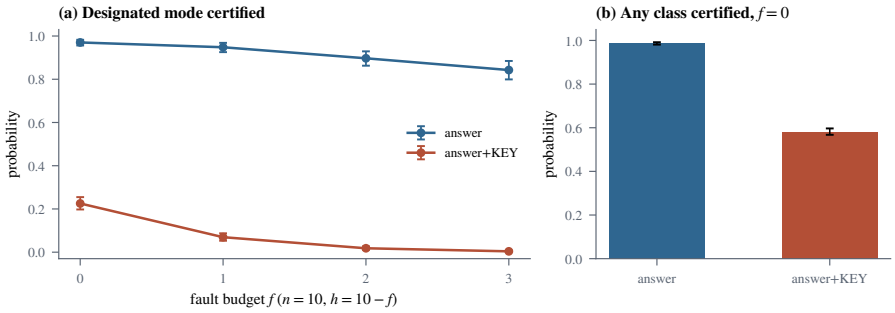

Figure 3: **Q2: dispersion, not the fault count, sets the certifiable budget.** (a) Probability that the designated empirical mode survives every append attack (honest margin $> 2f$), $n = 10$, $h = 10 - f$, three models averaged within task, 95% task-bootstrap intervals. (b) Probability that the rule certifies some class at $f = 0$.

C Scoring rules and regression tests

GSM8K: parse a decimal, strip currency, percent and digit-group separators, compare numerically. MMLU: extract the option letter from `B`, `B.`, `B)`, `**B**`. MBPP: parse the return value as a Python literal and compare typed structures recursively; `(0,6)` and `(0, 6)` coincide, `1` and `(1,)` do not. KEY: strip schema and markdown residue, collapse whitespace, casefold. Regression cases: `('Not matched!')` vs `Not matched!` same; `[1, 2]` vs `(1, 2)` different; `160 minutes`, `$160`, `160.0` vs `160` same; unparseable answer missing.

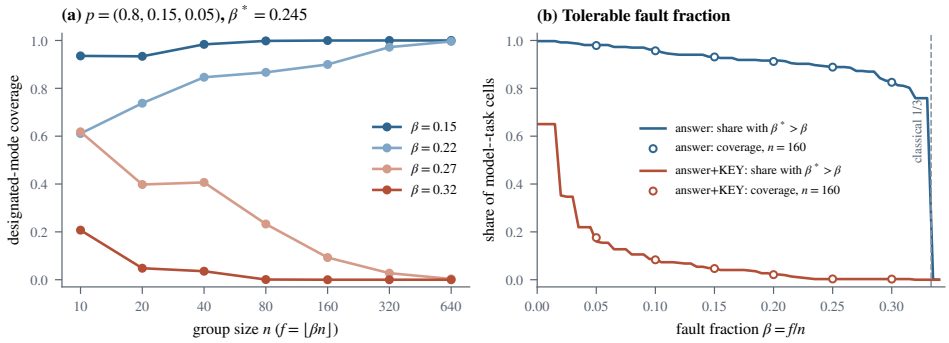

Figure 4: **Dispersion-limited resilience (Theorem 7).** (a) One law, $p = (0.8, 0.15, 0.05)$, $\beta^* = 0.245$: with $f = \lfloor \beta n \rfloor$, designated-mode coverage tends to one below β^* and to zero above it (Monte Carlo, 4,000 draws). (b) Share of the 369 model–task cells with $\beta^* > \beta$ (lines) and mean coverage at $n = 160$ (circles), answer and answer+KEY grains; the dashed line is the classical $1/3$.

D Dispersion curves and estimators

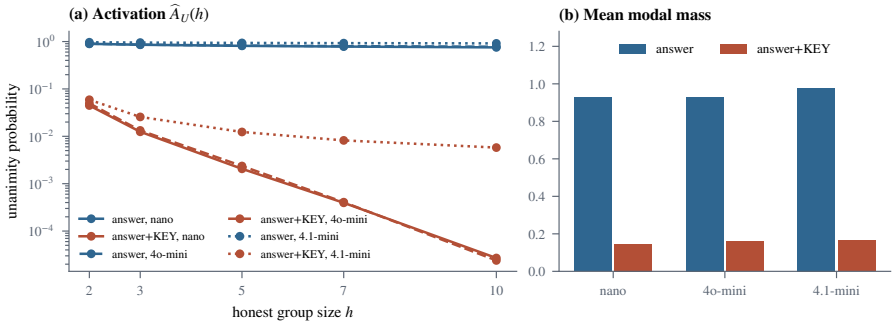

Figure 5: **Q1: finer agreement exposes dispersed proposals.** (a) $\widehat{A}_U(h)$ across group sizes; solid: answer, dashed: answer+KEY (lexical). (b) Mean modal mass. Common 123-task set.

Estimators. Plug-in vs U-statistic for `gpt-4.1-nano`: answer $h = 3$: 0.859 vs 0.855; $h = 10$: 0.767 vs 0.753; answer+KEY $h = 3$: 0.0167 vs 0.0125; $h = 10$: 1.43×10^{-4} vs 2.7×10^{-5} (ratio 5.2 at full precision).

E Further vendors

Four further models from three vendors at reduced sample size (Table 3). The finer-partition collapse replicates for `deepseek-v3.2`, `glm-4.7` and `qwen3-14b`. `qwen3-8b` returned near-identical replies at the nominal temperature (on 115 of 150 tasks all 16 replies share their first 80 characters; 2.3 distinct KEYs per task against 9.5–13.3 for the other three models), so it is reported as a near-deterministic sampler rather than as a representative of its family; its high $\widehat{A}_U(3)$ on the finer partition coincides with the lowest accuracy, the regime in which a high antecedent says nothing about correctness.

F Interaction, dissent and checker details

Interaction. Paired gains (round 2 minus round 0, percentage points, 95% task-bootstrap): $n = 3$: agreement 2.4 $[-2.0, 7.2]$, accuracy 0.0 $[-3.6, 4.4]$; $n = 5$: 4.8 $[-1.2, 10.8]$, 0.8 $[-2.8, 4.4]$; $n = 7$: 3.6 $[-2.8, 10.8]$, 2.4 $[-0.8, 6.4]$; $n = 10$: 7.6 $[0.8, 14.4]$, 1.6 $[0.0, 4.4]$.

Table 3: Further vendors at $k = 16$. The partition ordering replicates and answer-grain activation stays inside the main range. †Near-deterministic sampler at the nominal temperature.

model	tasks	accuracy	$\widehat{A}_U(3)$ answer	$\widehat{A}_U(3)$ answer+KEY	$p_{\max}$ answer+KEY
`deepseek-v3.2`	148	0.909	0.901	0.013	0.193
`glm-4.7`	150	0.916	0.904	0.013	0.200
`qwen3-14b`	150	0.839	0.882	0.076	0.320
`qwen3-8b`†	150	0.800	0.927	0.605	0.794

Mixed group (§6.1). slots 1–3 `gpt-4.1-nano`, 4–6 `gpt-4o-mini`, 7–10 `gpt-4.1-mini`; the same 50 GSM8K tasks, five repetitions, round 0 only; 3 of 250 debates had an unparseable reply and are excluded. Prediction $\sum_c \prod_m [n_{m,c}]_{k_m} / [N_m]_{k_m}$ from the stored single-agent replies, unbiased under independence. Round-0 majority accuracy (the plurality answer is correct) is 0.931, against 0.916 for homogeneous nano at $n = 10$.

Table 4: Parameter-free prediction of pre-interaction unanimity (`gpt-4.1-nano`, 50 GSM8K tasks). Predictions use single-agent histograms on the same tasks.

n	3	5	7	10
predicted $\widehat{A}_U(n)$	0.873	0.822	0.786	0.748
observed round-0 unanimity	0.868	0.812	0.792	0.756
observed − predicted (pp)	−0.5	−1.0	+0.6	+0.8

Dissent. Prompted dissent on `gpt-4.1-nano`: $n = 10$, three rounds, 150 planned trials per $f \in \{0, \ldots, 4\}$, of which 143–147 are retained after excluding unparseable honest replies; in round 2 attackers return the gold answer in 0.19–0.24 of slots (archived frame, 150 trials per f) and the assigned target in 0.74–0.78. Forced dissent on the same tasks: $f \in \{1, 2, 3\}$, three repetitions, the attacker's message fixed to the target without a model call. Forced vs prompted on `claude-opus-5`: Table 5. The 288 none/forced debates were archived first; the 144 prompted debates were collected later on the same task and configuration grid, so collection dates and backend snapshots are not controlled.

Table 5: Attack arms, round 2. Upper block: `claude-opus-5` on a common task/configuration grid (24 tasks, eight per benchmark, two repetitions), 48 trials per cell. Lower block: `gpt-4.1-nano` on the 50 GSM8K tasks of Table 4, three repetitions, trials with an unparseable honest reply excluded. "Target / slots" counts answers matching the designated wrong target among assigned attacker slots. None has zero assigned attackers; forced and prompted each assign f slots, whose realised messages define the measured attack.

arm	nominal f	trials	all unanimity	honest unanimity	honest acc.	target / slots
none	1	48	1.000	1.000	0.958	–
none	2	48	0.979	0.979	0.958	–
none	3	48	0.979	0.979	0.958	–
forced	1	48	0.000	1.000	0.958	48/48
forced	2	48	0.000	0.958	0.958	96/96
forced	3	48	0.000	0.979	0.958	144/144
prompted	1	48	0.667	1.000	0.958	0/48
prompted	2	48	0.625	1.000	0.958	1/96
prompted	3	48	0.438	0.958	0.958	9/144
nano forced	1	146	0.021	0.870	0.938	146/146
nano forced	2	144	0.014	0.875	0.924	288/288
nano forced	3	145	0.021	0.855	0.931	435/435
nano prompted	1	145	0.200	0.828	0.952	113/145
nano prompted	2	145	0.131	0.890	0.931	226/290
nano prompted	3	144	0.083	0.889	0.944	318/432

External checker case study. `claude-opus-5`, $n = 10$: 300 group trials find an accepted candidate in 0.900–0.917 of trials and none of 600 constructed wrong-return submissions passes, with up to four of ten slots assigned to attackers. A pooled acceptance rate of 0.80 predicts an empty pool with probability 10^{-7}; at $f = 0$ ($h = 10$), per-task rates predict 0.090, observed 0.083. Setup: 30 MBPP tasks, two repetitions per $f \in \{0, \ldots, 4\}$, candidate functions executed on the archived call and compared to the gold return; wrong-return attacks are wrapped candidates and their rejection is not a worst-case soundness estimate; the run evaluates a shared candidate pool and does not validate dissemination or safe execution of untrusted code. For $f = 0, 1, 2, 3, 4$, the honest pool is $h = 10, 9, 8, 7, 6$; abstention rates are $0.083, 0.100, 0.083, 0.083, 0.100$, each over 60 group trials.

G HUMAN CALIBRATION

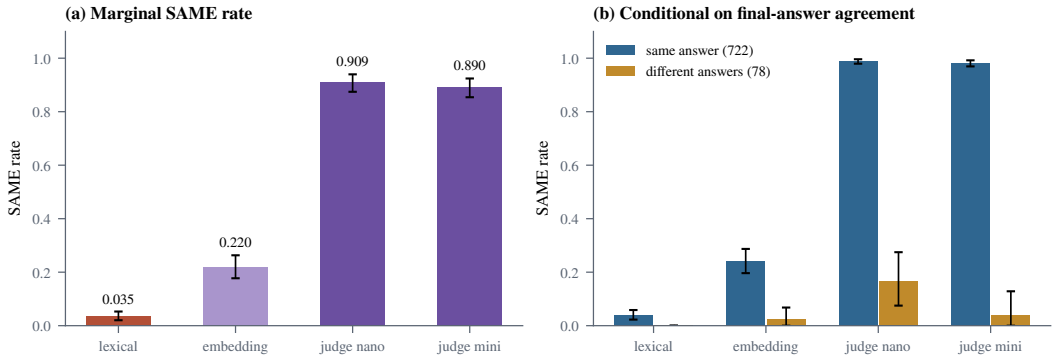

Figure 6: **Q3: different instruments report different agreement on identical pairs.** (a) Marginal SAME rates. (b) SAME rates conditional on final-answer agreement. 800-pair frame (722 same-answer, 78 different-answer under the current answer normaliser); observed relation frequencies, not human-calibrated accuracies. Error bars are 95% task-cluster percentile bootstrap intervals (5,000 resamples).

Design. Frame: all 518,837 unordered within-task pairs of stored chain-of-thought (CoT) replies from the three main models, including cross-model pairs; annotators saw the question and each reply's final answer and KEY, not its reasoning. Four disjoint strata, checked in order: S4 one or both KEYs missing (6,325 pairs); S2 different answer class (56,219); S3 same answer and identical normalised KEY (10,143); S1 same answer and different normalised KEY (446,150). Allocation: a 25-pair floor per stratum and the remainder proportional to $\sqrt{N_s}$, rounded to 300 by largest remainder, giving $148/69/43/40$ pairs for S1/S2/S3/S4; each pair carries weight N_s/n_s, so estimates target the pair frame, not the task-equal model-specific statistics of the main text. The sample covers 122 tasks and contains 113 same-model and 187 cross-model pairs. Three annotators labelled a 20-pair synthetic practice set, then the 300 pairs, on two axes: answer agreement (SAME/DIFF/UNSURE) and decisive-operation agreement (SAME/DIFF/VAGUE/UNSURE), where wording, language and symbolic form never constitute a difference. Reserved quantities: Fleiss' κ per axis; resolved joint-reference pairs (a $2/3$ majority on both axes, positive iff both SAME); weighted error, false acceptance and false rejection of each instrument against the joint majority; 95% within-stratum pair-bootstrap intervals (2,000 resamples). VAGUE, UNSURE and no-majority pairs are unresolved: excluded from error denominators and never counted as DIFF.

Results. 204 of 300 pairs resolve on both axes (weighted coverage $0.760\,[0.698, 0.816]$). Unweighted lexical confusion counts on resolved pairs: TP= 43, FP= 0, FN= 99, TN= 62 (unweighted false rejection 0.697; weighted 0.967 because S1, where the lexical rule always rejects, is 86% of the frame).

Pairwise collision estimates. Table 10 reports weighted SAME frequencies on the calibration frame: for a fixed partition these estimate pairwise collision $A(2)$; for the human reference they

Table 6: Human calibration. Zero observed false acceptance is not a population guarantee. Kappa is unweighted on the stratified sample; excluding any pair with an UNSURE operation rating gives $\kappa = 0.791$ on 207 pairs, a selected subset.

quantity	estimate	95% interval
weighted reference coverage	0.760	[0.698,0.816]
answer-normaliser error (answer axis)	0.006	—
lexical answer+KEY error	0.751	[0.693,0.807]
lexical false rejection	0.967	[0.963,0.970]
lexical false acceptance	0.000	—
human SAME among resolved lexical rejections	0.771	[0.712,0.828]
S1 human SAME among resolved pairs	0.852	[0.788,0.913]
answer-axis Fleiss' κ	0.947	—
operation-axis Fleiss' κ	0.532	—

Table 7: Instruments scored pair by pair on the 204 resolved calibration pairs against the joint human reference (weighted; 95% within-stratum bootstrap, 2,000 resamples). Judge prompts and the embedding threshold are those of §6.3; the two variants add a chain-of-thought instruction or three worked examples. The last two columns condition on the human answer majority over all 300 pairs (224 SAME, 76 DIFF), not only the resolved ones. Unweighted agreement between each judge and answer equality: nano $194/204$, mini $202/204$.

instrument	error	false accept	false reject	same-model error	SAME|ans SAME	SAME|ans DIFF
lexical answer+KEY	0.751 [0.69,0.81]	0.000 [0.00,0.00]	0.967 [0.96,0.97]	0.765	0.023	0.000
embedding, cosine ≥ 0.9	0.563 [0.48,0.64]	0.009 [0.00,0.03]	0.722 [0.64,0.80]	0.502	0.184	0.015
judge `gpt-4.1-nano`	0.149 [0.09,0.21]	0.666 [0.54,0.76]	0.000 [0.00,0.00]	0.138	0.990	0.164
judge `gpt-4.1-mini`	0.138 [0.08,0.20]	0.583 [0.44,0.69]	0.010 [0.00,0.03]	0.133	0.977	0.000
judge nano, chain-of-thought prompt	0.195 [0.14,0.26]	0.842 [0.75,0.91]	0.010 [0.00,0.03]	0.148	0.983	0.604
judge nano, few-shot prompt	0.153 [0.10,0.21]	0.685 [0.57,0.77]	0.000 [0.00,0.00]	0.143	0.992	0.197
answer equality only	0.132 [0.08,0.19]	0.583 [0.44,0.68]	0.003 [0.00,0.01]	0.138	0.993	0.000

measure pairwise answer+operation agreement without constructing a partition. The human range assigns unresolved pairs DIFFERENT or SAME, respectively, rather than giving a confidence interval. These are pair-weighted frame estimates, not the task-equal model-specific U-statistics of Table 1, and they do not recover a partition, $p_{\max}$ or $A(h)$ for $h \geq 3$; they do bound certification coverage, as follows.

Coverage bounds at the decisive-operation grain. Assume the human answer+operation relation is an equivalence relation that refines the answer partition (one of the 300 calibration pairs violates the refinement). For a model–task law r at that grain with modal mass $r_{\max}$ and answer-grain modal mass $p_{\max}$, $A_r(2) = \sum_c r(c)^2 \leq r_{\max} \sum_c r(c) = r_{\max}$, $r_{\max}^2 \leq A_r(2)$ and $r_{\max} \leq p_{\max}$, so $A_r(2)^h \leq r_{\max}^h \leq p_{\max}^{h-2} A_r(2)$. The upper side is linear in $A_r(2)$ and needs no homogeneity or independence assumption; the lower side uses $\mathbb{E}[A_r(2)^h] \geq (\mathbb{E}A_r(2))^h$. $A_r(2)$ is the plug-in collision probability, $1/k + (1 - 1/k)$ times the probability that two distinct replies are human-equivalent, so the bounds refer to the plug-in coverage of Table 2. It is estimated from the 102 same-model calibration pairs on the common tasks, each weighted by its stratum weight divided by the number of same-model frame pairs in its cell, so that every model–task cell counts equally. Unresolved pairs count as SAME for the upper bound and DIFFERENT for the lower bound. Intervals are within-stratum pair-bootstrap limits (2,000 resamples). The homogeneous column assumes that within every answer class the human relation has the same collision probability q, estimated as 0.686–0.885 on same-model same-answer pairs; then $r_{\max} \leq p_{\max}\sqrt{q}$ and coverage is at most $\mathbb{E}[p_{\max}^h]\, q^{h/2}$. As a check, the same estimator applied to answer equality gives $[0.673, 0.859]$ at $h = 7$, against the exact 0.843.

Decisive-operation partitions. Three annotators grouped 12 replies for each of 24 selected model–task cells (eight per benchmark and eight per model; two or three per benchmark–model pair), selected by stratifying on answer-grain modal mass. As in the pairwise calibration, they worked from blind sheets with model identities removed and a written guide; independence is not documented beyond the returned sheets. Annotators saw each reply's final answer and KEY, not its full reasoning. Replies placed in the same group by at least two annotators were linked, and connected components were intersected with normalised answer classes to form the answer+operation

Table 8: Designated-mode coverage at $n = 3f+1$ ($h = 2f+1$) under the human answer+operation relation: bounds from same-model calibration pairs (95% outer limits in brackets), the homogeneous within-answer-class estimate, and the exact answer-grain value.

f	h	lower bound	upper bound	homogeneous	answer grain
1	3	0.29 [0.17]	0.83 [0.91]	0.51–0.74	0.889
2	5	0.13 [0.05]	0.81 [0.88]	0.34–0.63	0.861
3	7	0.06 [0.02]	0.79 [0.87]	0.23–0.55	0.843

partition. The intersection was added after annotation, because the guide allowed replies with different answers to share an operation group (73 such pairs in 4 cells), and the closure adds 27 links in 5 cells that no two annotators made directly. VAGUE ratings are singletons in each annotator's partition. Coverage at $n = 3f + 1$ is estimated per cell by the U-statistic $[n_{c^*}]_h/[12]_h$ for the empirical mode c^*, which unlike the plug-in $(n_{c^*}/12)^h$ does not exceed $\widehat{A}_U(h)$; the plug-in overstates it by about 0.05 at $f = 3$. Intervals resample cells and, within each cell, the three annotators (2,000 replicates); they condition on the sampled replies. Mean operation-partition ARI is 0.523. The same 24 cells give 0.82 at the answer grain on their full reply pools and 0.0006 lexically; the lexical 0.004 of the main text is the mean over all 369 cells.

Table 9: Designated-mode coverage at $n = 3f + 1$ under three partitions of the same 288 annotated replies (unbiased estimates; last column plug-in). Brackets: 95% intervals at $f = 3$, over cells and annotators for answer+operation and over cells otherwise.

partition	$r_{\max}$	$\widehat{A}_U(3)$	$f = 1$	$f = 2$	$f = 3$	plug-in $f = 3$
answer	0.948	0.889	0.888	0.850	0.820 [0.688,0.951]	0.843
answer+operation	0.809	0.600	0.598	0.491	0.426 [0.271,0.607]	0.477
answer+KEY	0.201	0.0148	0.0146	0.0015	0.00005 [0,0.00016]	0.0014

Table 10: Weighted pairwise collision probabilities $A(2)$ on the calibration frame. "Human answer" is the human answer-axis majority; the answer-equality rule gives 0.888, 0.948, 0.853, 0.894, 0.961, 0.776 on the same rows. Human operation-axis SAME rates are 0.66–0.89 given the same human answer and 0.18–0.52 given different answers (bounds over unresolved pairs).

subset	pairs	human answer	lexical answer+KEY	human answer+operation	unresolved
all	300	0.895	0.020	[0.591, 0.831]	0.240
same-model	113	0.952	0.043	[0.667, 0.849]	0.182
cross-model	187	0.860	0.007	[0.545, 0.819]	0.274
GSM8K	123	0.894	0.011	[0.571, 0.792]	0.221
MBPP	77	0.980	0.021	[0.729, 0.908]	0.179
MMLU	100	0.776	0.036	[0.434, 0.796]	0.362

H FURTHER LIMITATIONS

Premise-versus-conclusion judgment aggregation (List & Pettit, 2002) is a further route to defining operation-level consensus. The certificate assumes a common view; the checker case assumes local verification, eventual delivery and one correct class per task. Interaction uses one model, three recorded rounds and no stopping rule; the opus attack arms were collected at different times. The ten-word KEY request is a design choice, and certification simulations condition on empirical laws from about 30 replies.

I REPORTING CHECKLIST: INSTANCES AND CONSEQUENCES

item	instance in this paper	consequence of omitting it
(1) grain and instrument	answer vs. answer+KEY; normaliser, lexical rule, embedding, judges	the same replies give coverage 0.843 or 0.004
(2) $A(h)$ by U-statistic	$\widehat{A}_U$ throughout; plug-in shown in Appendix D	plug-in over-states $A(10)$ by $5.2\times$ at the lexical grain
(3) calibrated label source	300-pair human calibration; error direction per instrument	lexical over-splitting (false rejection 0.967) or judge over-merging (false acceptance 0.58–0.67) is read as dispersion or agreement
(4) realised adversary	forced vs. prompted; adoption rate	a prompted opus attacker adopts 3.5% of its targets, so "tolerated f attackers" overstates the test
(5) dispersion vs. fault count	Remark 6; Theorem 7	at $n = 3f + 1$ dispersion alone makes the budget fail in 16% of draws at the answer grain
(6) accuracy beside agreement	Table 1; interaction gains	two debate rounds raise agreement 7.6 points and accuracy 1.6

J CLOSEST PRIOR RESULTS

Nearby impossibility results, and why we are not in them. The Consensus Trap (Liu et al., 2026) constrains anonymous outcome-level aggregators by symmetry; the Reasoning Trap (Shin, 2026) bounds an information quantity under closed-system debate; information-theoretic limits of safety verification (Scrivens, 2026) are detection-side. None concerns the validity antecedent. Bhardwaj et al. (2026) measure correlated failure across instances; Proposition 3 records the one interaction with mixtures. Relative to Berdoz et al. (2026), who report that benign agreement worsens with n, and Anand & Pappas (2026), who find benign consensus failures, we supply a mechanism consistent with them ($A(n)$ decays) and a parameter-free round-0 prediction; relative to Byzantine-agent evaluations (Berdoz et al., 2026; Anand & Pappas, 2026; Lee et al., 2026; Zheng et al., 2026), we add the attacker's adoption rate and a forced reference.

Table 11: Which elements each line of work makes explicit.

work	stochastic proposals	explicit grain	validity antecedent	certification margin	instrument calibration
self-consistency (Wang et al., 2023)	✓	answer only	–	–	–
semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024)	✓	✓	–	–	partial
agent BFT (Berdoz et al., 2026; Anand & Pappas, 2026; Lee et al., 2026; Zheng et al., 2026)	✓	implicit	–	$n/3$ or graph robustness	–
Xu et al. (2026)	✓	✓	–	observed f	instrument comparison
this paper	✓	parameter g_x	$A(h)$	honest $2f$; $\beta^* = \Delta/(2+\Delta)$	human-anchored

K CODE AVAILABILITY

The code and experimental artifacts are provided as supplementary material. They comprise the archived model replies and interaction transcripts, task records, annotation sheets, analysis scripts, and derived result files. The top-level `README.md` and `experiments/README.md` describe the experiment folders and identify the inputs and outputs for each analysis. Offline analyses can be inspected using the archived data; collecting new model replies requires the corresponding API access.