OUT OF CHARACTER: PER-AGENT CONFORMAL SCREENING FOR BYZANTINE LLM COMMITTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

Committees of language-model agents rely on redundancy to catch reasoning errors, but correlated exploits such as prompt injections can subvert multiple models simultaneously; when corrupted agents gain a majority, standard majority voting collapses from an uncorrupted accuracy of 0.705 to 0.094. Without labels, no confidence statistic of a single decision exposes an adversary that picks a wrong action while preserving the agent's confidence profile, and peer-consensus defences fail because the adversary forms the majority. We resolve this dilemma with **Sentinel**, a peer-free screen that tests each agent's confidence and action choices over a window against its own clean history, bounding the probability of falsely rejecting an honest agent by a budget δ in finite samples under exchangeability, regardless of how many peers are corrupted. Across 5 models and 3 benchmarks, Sentinel recovers 0.682 accuracy at 60% corruption and 0.650 at 80% within the screened window; on free-form answers under real prompt injection it regains 73–76% of the lost accuracy; and it restores live task completion from 0 to 12 of 165 tasks under a destructive attack; live under the exact null, honest rejection is 0.021–0.029 per screen with no attack, against a 0.10 budget. We further prove that an adversary preserving its own action law can change the *action symbol* on only a bounded fraction of decisions — a bound on symbol alteration, not on task failure.

1 INTRODUCTION

Committees of large language model (LLM) agents answer a question by pooling proposals through voting, debate or an automated judge (Du et al., 2024; Wang et al., 2025a; Zheng et al., 2023; Liang et al., 2024; Chan et al., 2024; Hong et al., 2024; Qian et al., 2024). Redundancy removes independent errors, not correlated ones: members sharing pre-training corpora, prompt templates and tools

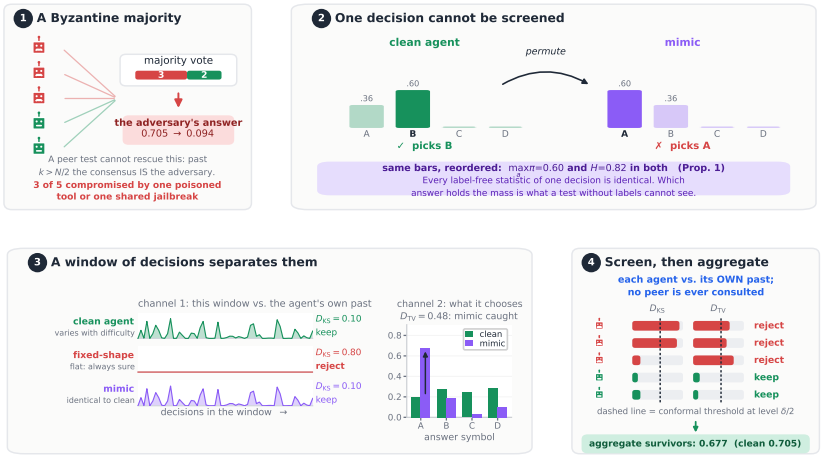


Figure 1: **The argument in four stages.** A Byzantine majority owns the vote, so the consensus *is* the adversary. One decision cannot be screened without labels: the mimic emits a *permutation* of the clean distribution. A *window* can be, on two label-free channels, the second catching the mimic. Panels (2) and (3) use cached Qwen2.5-7B outputs on MMLU.

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.

fail together, and one transferable jailbreak, prompt injection or poisoned retrieval artifact can subvert several at once (Greshake et al., 2023; Debenedetti et al., 2024; Zou et al., 2023; Huang et al., 2025). That is the Byzantine failure model (Lamport et al., 1982): once compromised members outnumber honest ones, every rule that counts votes returns the adversary's answer, and Byzantine-robust aggregators inherit the premise, since their median or distance-based selection follows the majority, which is now the attack.

Screening the members needs no labels, and the two obvious ways fail: a single-decision monitor sees only the shape of an output distribution, which an adversary can permute while acting wrongly (Proposition 1), and peer comparison inverts once the honest members are the minority; a trusted judge is a single point of failure the same injection reaches. Sentinel instead tests each member against its own recorded clean past, over a window, on two label-free channels: how confident it is, and what it picks. Exact permutation tests bound honest rejection in finite samples whatever the agent's outputs, and with no peer in the test that bound holds for every honest member whose decisions stay exchangeable with its calibration, however many peers are compromised (Theorem 2); survivors are aggregated by conformal trust weighting, which preserves each agent's marginal coverage (Proposition 4). The blind spot is exact: an adversary that redraws its action from its own clean symbol law leaves both channels calibrated, and how often it can change its action is capped by how predictable the agent was (Theorem 3).

Contributions.

- *Screening whose guarantee is independent of the corrupt count.* Exact permutation p-values against an agent's own clean history bound honest rejection in finite samples, with no labels, judge or peer comparison (Theorem 2); to our knowledge no prior defence for LLM committees keeps its validity under a corrupted majority.
- *Both directions of the detectability boundary.* No relabelling-invariant statistic of one decision separates an honest agent from a permutation of itself (Proposition 1); over a window an adversary is invisible to marginal screening exactly when it preserves the action law of its clean self, and can then change the action symbol on at most a share of decisions set by the concentration of that law (Theorem 3).
- *Recovery past a Byzantine majority, on benchmarks and live.* Across model families, benchmarks, attacks built against the screen itself and free-form answers under real prompt injection, the screened committee regains most of the accuracy a corrupted majority destroys, where robust aggregators, coordination protocols, peer screening and an LLM inspector do not, and the same screen runs inside a live tool-calling agent within its false-alarm budget.

2 Related work

Debate, mixture-of-agents, self-consistency and automated judging aggregate several model outputs into one prediction (Du et al., 2024; Liang et al., 2024; Wang et al., 2025a; 2023; Zheng et al., 2023; Chan et al., 2024; Hong et al., 2024; Qian et al., 2024; Yao et al., 2023), assuming members err independently. Transferable jailbreaks and prompt injections break that assumption by subverting several at once (Greshake et al., 2023; Debenedetti et al., 2024; Zou et al., 2023).

Byzantine-robust aggregation — coordinate median, trimmed mean, Krum and Bulyan (Blanchard et al., 2017; Yin et al., 2018; Chen et al., 2017; El Mhamdi et al., 2018) — filters outlying update vectors, but needs an honest majority: past it the median tracks the adversary and distance-based selection picks the attacking cluster. Such defences also degrade against adaptive attackers (Fang et al., 2020; Shejwalkar & Houmansadr, 2021), which motivates the evasion-optimal adversaries of §4.5.

Defences designed for LLM committees add topology filtering, external verifiers, staking and confidence-weighted consensus (Huang et al., 2025; Yu et al., 2025; Wang et al., 2025b; Chen et al., 2024a;b; Zheng et al., 2026; Lee et al., 2026; Jo & Park, 2025). They rest on peer inspection, a communication graph or a trusted judge: under majority corruption peer inspection inverts, a judge drawn from the committee is corruptible, and none bounds honest false rejection in finite samples. Receiver-side filtering (Lee et al., 2026) comes closest, but is guaranteed by a condition on the communication graph, not a false-rejection bound. Sentinel uses neither judge nor peer, and bounds that rate.

Split conformal prediction gives distribution-free finite-sample coverage under exchangeability (Vovk et al., 2005; Lei et al., 2018; Angelopoulos & Bates, 2023) and has been used to calibrate single LLMs (Kumar et al., 2023; Quach et al., 2024; Mohri & Hashimoto, 2024; Ren et al., 2023); conformal testing detects shift through exchangeable p-values (Vovk et al., 2003; Fedorova et al., 2012; Laxhammar & Falkman, 2015). We apply it to a committee, using each agent's own clean history as its null. Anomaly and backdoor detection likewise judge isolated inputs by confidence or representation geometry (Hendrycks & Gimpel, 2017; Lee et al., 2018; Yang et al., 2024; Gu et al., 2017; Tran et al., 2018; Xu et al., 2024; Hubinger et al., 2024), which Proposition 1 shows cannot separate an adversary that leaves single-decision statistics intact.

3 Motivating Case Study

Five LLM agents resolve customer requests on τ-bench with tool calls, and one shared prompt injection (Greshake et al., 2023; Debenedetti et al., 2024) compromises three of them, which we instantiate as the coordinated mimic of §5.1, keeping each member's confidence exactly. With no labels and the compromised members holding every vote, majority rule, debate and mixture-of-agents follow them, and coordinate median or Krum returns the attacking cluster. Peer screening, the principle of peer-inspection defences (Chen et al., 2024a; Wang et al., 2025b), inverts: its detection falls from 0.95 to 0.12 as the honest members become the minority (Figure 2(a)). A nonparametric CUSUM (Page, 1954; Ross & Adams, 2012) misses attacks that keep confidence intact, and an LLM inspector (Huang et al., 2025) is weaker still and reachable by the same injection. Sentinel compares each member with its own clean history, so its detection does not move with the corrupt count and Theorem 2 bounds honest rejection: screened, the committee acts like the uncorrupted one instead of executing the adversary's action (Figure 2(b)), and run live it restores completed tasks within budget (§5.8).

4 Method

Sentinel screens each agent against its own recorded clean history and aggregates the survivors. §4.2 shows why one decision cannot be screened (Proposition 1); §4.3 builds the window screen and its finite-sample guarantee, which holds for any number of corrupted agents (Theorem 2); §4.4 aggregates; §4.5 characterises the adversaries no marginal screen can see (Theorem 3). Figure 1 illustrates the first two steps, and every result points to its proof in Appendix B.

4.1 Setup and Notation

Agents and decisions. A *decision point* x has a finite *candidate set* $\mathcal{A}(x)$ whose size may vary, such as the options of a question or an agent's available tool calls. A *committee* is N agents indexed by $i \in [N] = \{1, \dots, N\}$. Agent i returns a distribution $\pi_i(\cdot \mid x)$ in the probability simplex $\Delta(\mathcal{A}(x))$: a local model's softmax over the candidate labels, or an API model's renormalised log-probabilities. The method observes nothing else; free-form generation is handled in §5.9. An aggregation rule maps the N distributions to one action $\hat{a}(x)$, and $a^\star(x)$ is the correct action, which is never available at test time.

Label-free statistics. For a distribution π over candidates, the *confidence complement* and the *entropy* are

$$q(\pi) = 1 - \max_a \pi(a), \qquad H(\pi) = -\sum_a \pi(a) \log \pi(a), \tag{1}$$

and we write $q_i(x) = q(\pi_i(\cdot \mid x))$ and $H_i(x) = H(\pi_i(\cdot \mid x))$.

Symbols. Candidate indices are not comparable across decisions, so a *symbol map* $\sigma : \bigcup_x \mathcal{A}(x) \to \mathcal{S}$ sends each action to a stable label in a finite alphabet $\mathcal{S}$ of size V: the answer letter, the command verb, the tool name. Agent i's *selected symbol* is $\hat{\sigma}_i(x) = \sigma(\arg\max_a \pi_i(a \mid x))$, and its *$\sigma$-marginal* P_i is the law of $\hat{\sigma}_i(x)$ over clean decisions.

Threat model. An adversary controls a fixed set $\mathcal{C} \subset [N]$ of $k = |\mathcal{C}|$ agents for a session and may replace their outputs by *any* distributions, coordinated across $\mathcal{C}$ and chosen with full knowledge of the clean committee (Lamport et al., 1982). The defender knows neither $\mathcal{C}$ nor k, observes only $\{\pi_i(\cdot \mid x)\}_{i \in [N]}$, and has no labels at test time. We target $k > N/2$, where every vote-counting rule fails. An attack is *fixed-shape* if it replaces a corrupted agent's output by a distribution of fixed form whatever the decision, such as near-certainty on a wrong candidate.

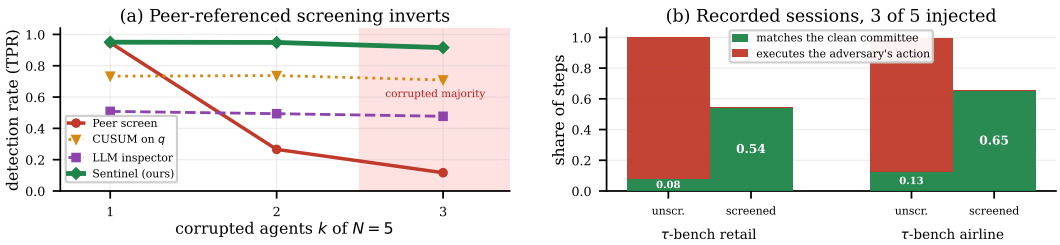


Figure 2: **(a)** Detection against the corrupt count, benchmark suite. **(b)** Replayed τ-bench sessions, 3 of 5 injected: share of steps matching the uncorrupted committee (green) or the adversary's target (red).

Calibration and budgets. A labelled set $\mathcal{D}_{\text{cal}}$ of n decision points is recorded while all agents are clean. The defender fixes a miscoverage level α for conformal sets and a *false-rejection budget* $\delta \in (0, 1)$, the largest probability with which the screen may reject a clean agent.

4.2 One decision is not enough

Without labels, a test of one decision can only examine the shape of an agent's output distribution, and an adversary that keeps the shape is invisible to it.

Proposition 1 (Per-decision indistinguishability)**.** *Let a corrupted agent emit* $\pi_i'(\cdot \mid x) = \pi_i(\varrho_x^{-1}(\cdot) \mid x)$ *for a permutation* ϱ_x *of* $\mathcal{A}(x)$*. Every statistic* ϕ *of a single output distribution that is invariant to relabelling the candidates satisfies* $\phi(\pi_i') = \phi(\pi_i)$*; in particular* q *and* H *are unchanged.* [Proof: Appendix B.1]

The *permutation mimic* exploits this: at every decision it moves its clean argmax mass onto a wrong candidate $a^\dagger$, so it is always wrong while its q and H remain those of its clean self. Two detectors fall outside the proposition, and both fail in our threat model. One that reads the decision's content is not relabelling-invariant but needs a trusted reader; we evaluate the strongest such reader, an LLM inspector, in §5.4. One that compares an agent with its peers fails once $k > N/2$, because the peer majority is the adversary. What remains is how an agent behaves across many decisions.

4.3 Screening over a window

A clean agent's confidence varies with the difficulty of each decision, whereas a fixed-shape adversary holds it constant; an adversary that steers the committee toward a target changes which symbols the agent selects. Neither shows at one decision, but both show over a window. Sentinel therefore screens a window $\mathcal{W}$ of W consecutive decisions before aggregating them, comparing each agent with its own clean history on two label-free channels; the committee is unscreened until the first window closes (Appendix E.7).

Two channels. Let Q_i^{cal} and P_i^{cal} be the empirical laws of $q_i(x)$ of (1) and of $\hat{\sigma}_i(x)$ over $\mathcal{D}_{\text{cal}}$, and $Q_i^{\mathcal{W}}$ and $P_i^{\mathcal{W}}$ the same laws over $\mathcal{W}$. With F_Q the distribution function of Q and $\|\cdot\|_{\text{TV}}$ the total variation distance, the channels are

$$D_i^{\text{KS}} = \sup_t \left| F_{Q_i^{\mathcal{W}}}(t) - F_{Q_i^{\text{cal}}}(t) \right|, \qquad D_i^{\text{TV}} = \left\| P_i^{\mathcal{W}} - P_i^{\text{cal}} \right\|_{\text{TV}}, \tag{2}$$

which detect a change in *how confident* the agent is and in *what it chooses*.

Calibrated rejection. Each channel is calibrated by permutation. Pool agent i's W window decisions with its n calibration decisions, split the pool uniformly at random B times into sets of sizes W and n, and recompute the channel on each split. For a channel $c \in \{\text{KS}, \text{TV}\}$ with observed value D_i^c and split values $D_{i,b}^c$, the p-value is

$$p_i^c = \frac{1 + |\{b \le B : D_{i,b}^c \ge D_i^c\}|}{B + 1}, \tag{3}$$

and agent i is *rejected* when (a null built from resampled windows instead is not exchangeable and overspends the budget, Appendix E.11)

$$\min\left(p_i^{\text{KS}}, p_i^{\text{TV}}\right) \le \delta/2. \tag{4}$$

Theorem 2 (Validity and consistency of the screen). *(i) If a clean agent's W window decisions and its n calibration decisions are exchangeable, rule (4) rejects it with probability at most δ, for every W, n and B and whatever its output distribution. (ii) If a corrupted agent's window decisions are i.i.d. with a confidence law or a σ-marginal different from its calibration laws, the rule rejects it with probability tending to one as $W, n \to \infty$, for any $B \geq 2/\delta$.* [Proof: Appendix B.2]

Part (i) involves no other agent, so no number of corrupted agents can loosen it; it holds per agent, and level δ/N bounds by δ the chance that any clean agent is rejected. Part (ii) covers every fixed-shape attack, since clean confidence varies across decisions, and every attack that selects some symbol more often than the clean agent does. Part (i) uses nothing about the statistic, so any channel computed alike on every split inherits it, continuous ones included (Proposition 7).

Dependent decisions. Consecutive decisions of an agent acting in an environment share a task and are not exchangeable, but whole episodes are. The window is then L whole episodes, with L fixed from calibration episode lengths to cover about W decisions, and the pool is split by episode; part (i) holds with episodes in place of decisions (§5.7).

Before the first window. An anytime-valid test covers the first W decisions. With $p_{i,t}$ the smoothed conformal p-value of q_i at decision t against calibration and earlier decisions, independent and uniform under exchangeability (Vovk et al., 2003), $E_{i,t} = \prod_{u \leq t} g(p_{i,u})$ is a mean-one nonnegative martingale for any betting density g on $[0, 1]$ fixed in advance, so by Ville's inequality (Ville, 1939) rejecting once $E_{i,t} \geq 2/\delta$ has false-rejection probability at most $\delta/2$ per session, and with the window screen at $\delta/2$ after W the total is at most δ. We do not evaluate this extension.

4.4 Aggregation over survivors

Let $S \subseteq [N]$ be the agents that rule (4) keeps, with $S = [N]$ if it rejects all of them (Appendix E.6). Any rule can aggregate S; we report majority vote and *conformal trust weighting*.

Conformal sets. Each agent keeps the set

$$C_i(x) = \{a \in \mathcal{A}(x) : \pi_i(a \mid x) \geq 1 - \hat{\eta}_i\}, \tag{5}$$

replaced by its argmax if empty, where $\hat{\eta}_i$ is the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile of the scores $1 - \pi_i(a^\star \mid x)$ over $\mathcal{D}_{\text{cal}}$. It contains $a^\star(x)$ with probability at least $1 - \alpha$ when x is exchangeable with the calibration points (Proposition 4, Appendix B.4).

Trust weights. Agent i's weight at x is

$$T_i(x) = \frac{\exp(-z_i(x))}{|C_i(x)|\,(H_i(x) + \epsilon)^\beta}, \qquad z_i(x) = \max\left(\left| \tfrac{H_i(x) - \mu_i^H}{\varsigma_i^H} \right|, \left| \tfrac{q_i(x) - \mu_i^q}{\varsigma_i^q} \right| \right), \tag{6}$$

where (μ_i^H, ς_i^H) and (μ_i^q, ς_i^q) are the calibration mean and standard deviation of H_i and q_i and $\epsilon > 0$ is a numerical floor. The weight shrinks with the set size, penalises departure from the agent's own calibration statistics, and rewards confidence through $\beta = \exp(-\bar{\mu}^H / (\kappa \log V))$, where $\bar{\mu}^H$ is the committee's mean calibration entropy and $\kappa = 0.15$: $\beta \to 1$ on peaked tasks and $\beta \to 0$ on diffuse ones. The committee's action is $\hat{a}(x) = \arg\max_a \sum_{i \in S} T_i(x)\, \pi_i(a \mid x)\, \mathbf{1}[a \in C_i(x)]$; Algorithm 1 (Appendix C) states the procedure end to end.

4.5 What a marginal screen cannot see

The permutation mimic is invisible to D^{KS} by construction and must be caught by D^{TV}. How well it hides depends on how it picks its target.

Mimics. A *session-scoped* mimic targets one symbol for a whole session, so its σ-marginal shifts with its success and it is rejected (Proposition 5, Appendix B.4). A *step-scoped* or *marginal* mimic draws its target afresh from P_i at every decision and leaves both channels clean: over 600 decisions it stays within 0.025 of the clean marginal in total variation, against the permutation mimic's 0.088.

Refining the alphabet. A map σ' *refines* σ if $\sigma = f \circ \sigma'$ for some f. An adversary that preserves a finer marginal preserves every coarser one (Proposition 6), so refinement narrows the escape. We use the ordered *couple*

$$\sigma^{(2)}(\pi) = \big(\sigma(\arg\max_a \pi(a)),\ \sigma(\arg\min_a \pi(a))\big) \in \mathcal{S}^2, \tag{7}$$

which records what an agent ranks last as well as first and raises the separation the marginal mimic must hide from 0.025 to 0.107, a factor of 4.3. The *union screen* adds the couple's total variation distance as a third channel and rejects when any of the three p-values is at most $\delta/3$, keeping Theorem 2(i).

Coupling adversaries. How far can refinement go? Assume every symbol in the support of P_i is available at every decision, as in multiple choice. A *coupling adversary* with Markov kernel K on $\mathcal{S}$ replaces the clean selected symbol $s = \hat{\sigma}_i(x)$ by $s' \sim K(s, \cdot)$, drawn independently across decisions, and permutes $\pi_i(\cdot \mid x)$ so that its argmax carries s'. The kernel is *stationary* if $\sum_s P_i(s)\, K(s, s') = P_i(s')$ for every s'; the marginal mimic is the stationary kernel $K(s, \cdot) = P_i$. The *alteration rate* is the probability that the adversary changes a decision's selected symbol.

Theorem 3 (The limit of marginal screening)**.** *Fix a symbol map with finite range and let* $m = \max_s P_i(s)$. *(i) If the clean window decisions are i.i.d., a coupling adversary with a stationary kernel leaves* q_i *and* H_i *unchanged at every decision; every test that depends on the window only through the empirical law of* q_i*, or only through the empirical* σ*-marginal, has power equal to its size against it; rule (4) rejects it with probability at most* δ*, and so does the union screen when the map is the couple. (ii) Over stationary kernels the largest alteration rate is* $\min\{1,\, 2(1-m)\}$*; for a refinement of the map, the largest rate at which the selected symbol changes is the value of a transportation programme and is no larger.* [Proof: Appendix B.3]

Two consequences follow. Evasion does not force incoherence: the maximising kernel maps each clean answer to a different one, so when no symbol holds more than half of P_i, as on balanced multiple choice, the adversary alters every decision and only a task-level metric or content reader catches it. Where one symbol dominates, the agent's habits cap the alteration rate at $2(1-m)$, 0.22 on the tool-use banks of §5.8, where one symbol holds 0.89 of decisions. The cap is on symbols, not harm, and assumes P_i preserved exactly; an adversary that estimates it is detectable to that extent (§5.6, §A).

Arguments. A tool-name map leaves arguments free; screening the pair (σ, ψ), with ψ a finite map of the arguments, caps argument tampering under a kept tool at $\sum_s P_i(s) \min\{1, 2(1-m_{i,s})\}$, attained, $m_{i,s}$ being the largest share of one argument class under tool s (Corollary 8).

5 Experiments

5.1 Setup

Committees and benchmarks. Unless stated otherwise $N = 5$ and the committee is *homogeneous*: five personas of one base model — Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, Phi-3.5-mini-instruct, Qwen3-30B-A3B-Instruct-2507 or OLMo-3-7B-Instruct — differing only by system prompt; heterogeneous and frontier committees are in §5.3. Benchmarks are MMLU (Hendrycks et al., 2021) (16 subjects), TruthfulQA (Lin et al., 2022) and ARC-Challenge (Clark et al., 2018), with GPQA-Diamond (Rein et al., 2024) as a hard regime (Appendix E.8). Agentic evaluation replays per-decision banks from ALFWorld (Shridhar et al., 2021) and Mind2Web (Deng et al., 2023) and runs live on τ-bench (Yao et al., 2025).

Attacks. Five *fixed-shape* attacks (overconfident, overconfident-extreme, coordinated-extreme, subtle, random) replace a corrupted agent's distribution by one of fixed form. Two *defence-aware* attacks follow Proposition 1: the *permutation mimic* moves the mass of the agent's own clean distribution onto a wrong action, preserving its confidence exactly, and the *coordinated mimic* does so toward a shared target; §5.6 adds those of §4.5.

Metrics. *Accuracy* is that of the committee's action. The screen's *TPR* is the fraction of corrupted agents it rejects and its *FPR* the fraction of clean agents it rejects, the quantity δ bounds; H_0 denotes $k = 0$. *AUROC* scores detection at every level at once, by the probability that a corrupted agent gets a smaller $\min(p_i^{\mathrm{KS}}, p_i^{\mathrm{TV}})$, the score rule (4) thresholds, than a clean one. *Recovery* is the share of the accuracy lost to an attack that a defence regains, $(\mathrm{acc} - \mathrm{acc}_{\mathrm{maj}})/(\mathrm{acc}_{\mathrm{clean}} - \mathrm{acc}_{\mathrm{maj}})$.

Protocol. $\alpha = \delta = 0.10$, $W = 200$, $B = 200$, $\kappa = 0.15$, 20 seeds. The corrupt set is redrawn per seed and fixed for the session; calibration and test splits are disjoint and reshuffled per seed. Intervals are 95% bootstrap intervals over seeds, and significance is a one-sided paired Wilcoxon signed-rank test over seeds against the strongest competitor, Holm-corrected across attacks.

5.2 Recovery past a Byzantine majority

At $k = 3$ of $N = 5$ (Table 1) majority vote collapses to 0.000 under the coordinated attack and 0.089–0.227 otherwise, against a clean 0.705, and aggregators built for an honest majority collapse with it (0.000–0.238): the coordinate median (Yin et al., 2018) becomes the adversary's value, and Krum (Blanchard et al., 2017), even given the true k, selects the tightest cluster, which is the

Table 1: Accuracy at $k = 3$ of $N = 5$, 5 models $\times$ 3 benchmarks $\times$ 20 seeds (Insp.: Huang et al. (2025), 10 seeds); clean 0.705. *Median* equals the trimmed mean; ReCon.: ReConcile; AUROC, TPR: the screen's detection; *Recov.*: share of lost accuracy regained. Sentinel: 95% half-width, paired Wilcoxon vs. the row's strongest competitor, Holm-corrected ($^{**}p < 0.01$). †: defence-aware.

	screen		accuracy							
Attack	AUROC	TPR	Maj.	Median	Krum	Bulyan	ReCon.	Insp.	**Sentinel**	Recov.
overconfident	0.994	0.93	0.089	0.040	0.006	0.143	0.141	0.633	**0.701**$_{\pm.006}$**	99%
overconfident-extreme	0.994	0.93	0.089	0.034	0.006	0.117	0.116	0.548	**0.693**$_{\pm.007}$**	98%
coordinated-extreme	0.994	0.93	0.000	0.000	0.000	0.000	0.000	0.601	**0.693**$_{\pm.007}$**	98%
subtle	0.994	0.93	0.143	0.080	0.126	0.704	0.697	**0.706**	**0.706**$_{\pm.005}$	100%
random	0.994	0.93	0.227	0.227	0.227	**0.706**	0.702	**0.706**	**0.706**$_{\pm.005}$	100%
coordinated-mimic†	0.994	0.93	0.024	0.034	0.219	0.036	0.024	0.384	**0.667**$_{\pm.015}$**	94%
permutation-mimic†	0.940	0.81	0.089	0.033	0.238	0.081	0.086	0.355	**0.609**$_{\pm.021}$**	84%
Mean	0.986	0.92	0.094	0.064	0.117	0.255	0.252	0.562	**0.682**$_{\pm.007}$	96%

Table 2: Committee composition, $k = 3$ of $N = 5$, seven attacks, 20 seeds. *Clean*: uncorrupted majority vote; Sentinel: 95% half-width; *Recov.*: share of lost accuracy regained.

Committee	Clean	Majority	TPR	FPR	**Sentinel**	Recov.
Homogeneous (5 base models)	0.705	0.094	0.92	0.04	**0.682**$_{\pm.007}$	96%
Heterogeneous (5 models)	0.774	0.208	0.94	0.06	**0.734**$_{\pm.028}$	92%
Heterogeneous + frontier member	0.810	0.228	0.96	0.06	**0.777**$_{\pm.014}$	94%
Frontier personas (Qwen3.8-27B)	0.880	0.126	0.96	0.03	**0.859**$_{\pm.021}$	97%

attack. The strongest prior defence, the LLM inspector of Huang et al. (2025), reaches 0.355–0.706 (Appendix E.4). Sentinel scores 0.667–0.706 on six of seven attacks and 0.609 on the permutation mimic, recovering 96% of the lost accuracy, significantly above the strongest competitor on five ($p < 0.01$) and tied on subtle and random, at AUROC 0.994 (0.940 on the mimic). No count can substitute (Figure 3(a)): majority vote is flat until it flips (0.705 to 0.094 from $k = 0$ to 3) while the screened committee declines only to 0.682, because the corrupt fraction never enters the test, and at $k = 4$ it scores 0.650 against 0.047 (Appendix E.5).

5.3 Heterogeneous and Frontier Committees

Judged against its own history, an unequal member is screened as well as a clone (Table 2, Figure 4(c)): across the four compositions Sentinel regains 92–97% of the lost accuracy at FPR 0.03–0.06, and a frontier member does not raise false rejection.

5.4 Screening Without Peers

Table 3 compares screens that need neither labels nor a trusted majority. A peer screen detects 0.95 of attackers at $k = 1$ but 0.27 and 0.12 at $k = 2$ and 3, since past $N/2$ the majority it consults is the adversary; a nonparametric CUSUM on confidence (Page, 1954; Ross & Adams, 2012) detects 0.71 and the LLM inspector 0.48 at $k = 3$. Sentinel detects 0.95, 0.95 and 0.92 at FPR 0.04, flat in k, and its accuracy of 0.682 is significantly above the inspector's 0.562 ($p < 0.001$, 10 paired seeds). The channels are complementary (Appendices E.2 and E.3): D^{KS} reaches TPR 0.97 on fixed-shape attacks but AUROC exactly 0.500 on both mimics, as Proposition 1 requires; D^{TV} catches the coordinated mimic (0.96) but misses attacks that move confidence, not choices (0.83 and accuracy 0.649 on overconfident-extreme, against the pair's 0.93 and 0.693); over 20 paired seeds the pair beats it by +0.018 TPR ($p = 0.023$) and +0.004 accuracy ($p = 0.045$). A CUSUM in place of D^{KS} spends more budget and forfeits Theorem 2, and unscreened trust weighting falls to 0.160, since a mimic's preserved confidence earns it trust.

Table 3: Screens at k of $N = 5$ corrupted, means over the seven attacks (inspector: 10 seeds). Bold: best in column. FPR and accuracy are at $k = 3$; Sentinel carries its 95% half-width and a paired Wilcoxon test against the best prior screen ($^{***}p < 0.001$). Channel ablations and budgets: Appendix E.2.

	TPR				
Screen	k=1	k=2	k=3	FPR	Accuracy
Peer screen	**0.95**	0.27	0.12	**0.04**	0.470
CUSUM on q	0.73	0.74	0.71	0.08	0.556
LLM inspector	0.51	0.49	0.48	0.09	0.562
Sentinel	**0.95**	**0.95**	**0.92**	**0.04**	**0.682**$_{\pm.007}$***

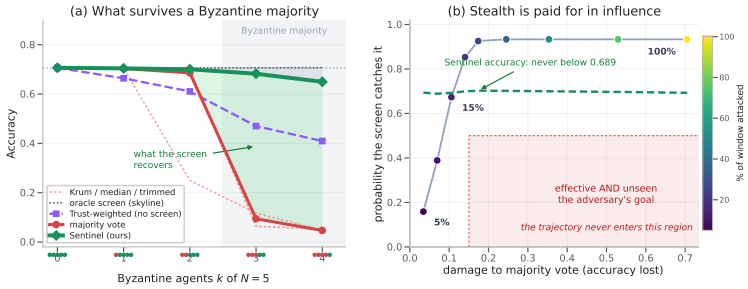


Figure 3: **(a)** Accuracy against corrupted agents k, over 5 models $\times$ 3 benchmarks $\times$ 7 attacks $\times$ 20 seeds; shaded: what the screen recovers over majority vote, hatched: the gap to a perfect screen. **(b)** Damage against probability of detection, coloured by the attacked share of the window.

Table 4: Coordination protocols as multi-round LLM calls; 200 questions, 12 seeds. MoA: Mixture-of-Agents. Sentinel: 95% half-width, paired Wilcoxon vs. the best protocol at each attacked k, Holm-corrected ($^{**}p < 0.01$). Last row: the worst seed at $k = 3$.

k	Majority	Debate	MoA	LLM-as-judge	Trust-weighted	**Sentinel**
0	0.754	0.713	0.743	0.727	0.759	**0.760**$_{\pm.012}$
1	0.752	0.694	0.621	0.601	0.711	**0.760**$_{\pm.011}$ **
2	0.736	0.684	0.309	0.309	0.654	**0.742**$_{\pm.015}$
3	0.070	0.071	0.353	0.381	0.466	**0.704**$_{\pm.033}$
worst seed	0.015	0.056	0.019	0.013	0.421	**0.556**

5.5 Coordination protocols

Table 4 runs debate (Du et al., 2024), Mixture-of-Agents (Wang et al., 2025a) and LLM-as-judge (Zheng et al., 2023) as multi-round LLM calls with a corruptible aggregator or judge. Mixture-of-Agents and LLM-as-judge collapse at $k = 2$ (0.309), since one corrupted aggregation point suffices, and debate with the vote at $k = 3$, where Sentinel scores 0.704 against the best protocol's 0.381. That mean hides a coin flip: on the 6 of 12 seeds where the judge is drawn honest it scores 0.746 against Sentinel's 0.712, and 0.016 against 0.697 on the rest, so the paired test at $k = 3$ is not significant while Sentinel's worst seed (0.556) is far above every protocol's (≤ 0.056).

5.6 Adaptive adversaries

Attacking part of the window. An adversary can attack a fraction of the window and leave the rest calibrated (Figure 3(b), Appendix E.9): at 5% detection falls from TPR 0.93 to 0.16 but majority vote still scores 0.671 of a clean 0.705, and once the attack halves the vote detection is back at 0.93, as for ramps. The screened committee never falls below 0.689: no rate is both hidden and effective.

Refining the symbol alphabet. The union screen of §4.5 gains +0.14 to +0.16 detection on the mimics, every 95% interval over 300 matched triples excluding zero (Appendix E.10).

The limit adversaries. Built as Theorem 3 prescribes (Table 8, Appendix E.1), the adversaries confirm the boundary. No letter carries half of any agent's clean answers, so the programme's value is 1.00; the two-channel screen is blind to all three constructions (TPR 0.04–0.07 at FPR 0.09), the union screen catches the one that breaks the couple law (TPR 0.68), and the couple-law coupling defeats both, holding the committee at 0.305 against majority vote's 0.260.

5.7 Agentic domains

Consecutive decisions share a task and are not exchangeable: on the ALFWorld and Mind2Web banks (Table 5) re-splitting single decisions rejects honest members at 0.14–0.24, up to two and a half times the budget, because the hypothesis of Theorem 2 fails; re-splitting whole episodes restores the budget in all ten cells (0.025–0.050) and makes detection rise with the window (Appendix E.12).

5.8 Live tool calling on τ-bench

On τ-bench a five-member committee converses with a user simulator and calls real tools (17 candidates in retail, 15 in airline). Replayed from recorded banks the screen detects attackers at TPR

Table 5: Agentic banks, TPR / FPR, re-splitting single decisions (i.i.d.) or whole episodes (block), $k = 3$ of 5, 60 seeds; stratified nulls: Table 21. Underlined: outside $\delta = 0.10$.

Domain	Null	W=16	W=32	W=64	W=128	W=256
ALFWorld	i.i.d.	0.833 / 0.208	0.683 / 0.200	0.583 / 0.183	0.617 / 0.233	0.533 / 0.200
	block	0.828 / 0.042	0.828 / 0.042	1.000 / 0.025	1.000 / 0.042	1.000 / 0.025
Mind2Web	i.i.d.	0.617 / 0.192	0.417 / 0.142	0.567 / 0.217	0.567 / 0.233	0.617 / 0.242
	block	0.950 / 0.042	1.000 / 0.033	1.000 / 0.050	1.000 / 0.042	1.000 / 0.033

Table 6: Live τ-bench, $k = 3$ of 5 personas, exact null, warm-up calibration (other sources: Table 25). FPR_0: honest rejection per call, no attack; FPR, TPR: per call, attacked; last columns: tasks completed, attacked. Bold: best per domain; underlined: outside $\delta = 0.10$.

			k=0	k=3		tasks completed	
Domain	Calibration	Attack	FPR_0	FPR	TPR	unscreened	screened
retail	warm-up, exact	mimic	**0.021**	**0.000**	**0.97**	2/30	**4/30**
	warm-up, exact	destructive	**0.021**	0.017	0.93	1/30	**4/30**
airline	warm-up, exact	mimic	**0.029**	**0.018**	**0.96**	1/30	**3/30**
	warm-up, exact	destructive	**0.029**	**0.018**	**0.96**	0/30	**3/30**

Table 7: Free-form GSM8K, 3 of 5 injected Qwen2.5-7B personas, 20 seeds. Sentinel: 95% half-width, paired Wilcoxon vs. best baseline ($^{***}p < 0.001$); honest rejection with no attack 0.01–0.03.

Injection	Clean	Majority	Peer	Per-decision	**Sentinel**	TPR / FPR
targeted	0.866	0.660	0.660	0.647	$\mathbf{0.810}_{\pm.017}{}^{***}$	1.00 / 0.025
stealthy	0.856	0.624	0.624	0.616	$\mathbf{0.797}_{\pm.020}{}^{***}$	1.00 / 0.075
subtle	0.867	0.644	0.644	0.643	$\mathbf{0.813}_{\pm.019}{}^{***}$	1.00 / 0.075

1.00 in both domains; a committee of five personas of one strong model (Appendix E.13) has one dominant symbol, which caps Theorem 3's coupling at 0.37–0.43 of its actions (Appendix E.14).

Live rollouts. Table 6 screens inside live rollouts at $k = 3$ of 5. Bank calibration is unsafe (honest rejection 0.144 on retail with no attack, 0.034 on airline; Appendix E.16); the committee's own clean warm-up holds in both (0.078, 0.058), and under the exact null of Theorem 2 rejection falls to 0.021 and 0.029 per screen and 0.033 per session, with detection 0.93–0.97 and more tasks completed in every row. Five different models exceed the budget under attack (Appendix E.15). Under the coordinated-extreme attack, over all 165 tasks, majority solves 0 against the screened committee's 12 and a clean 11 (McNemar $p = 0.0005$), at detection 0.69 and 0.93 and attacked FPR 0.048 and 0.116.

5.9 Open-ended generation under prompt injection

Five Qwen2.5-7B personas answer GSM8K (Cobbe et al., 2021) in free text, three with injected prompts (Appendix E.17). Sentinel's channels become KS on answers' mean log-probability and the energy distance (Székely & Rizzo, 2013) of their embeddings, permutation-calibrated (Proposition 7): it rejects every injected member at FPR below 0.08 and regains 73–76% of the lost accuracy (Table 7). The peer screen inverts, and a per-decision representation divergence catches only 0.12–0.15 of injected answers.

6 Conclusion

Voting and Byzantine-robust aggregation both trust the majority, so both fail once compromised agents hold it. Sentinel instead judges each agent against its own clean history: exact permutation tests bound honest rejection in finite samples whatever the number of compromised peers, for every agent that stays exchangeable with its calibration. We characterised exactly which adversaries a marginal screen cannot see and how much they can alter, and on multiple choice, injected free-form answers and live tool calling the screened committee regains most of the accuracy a corrupted majority destroys.

AI USE STATEMENT

Generative AI tools were used in the preparation of this paper for language editing and for drafting and revising prose under the authors' direction, and as a coding assistant for the analysis, plotting and verification scripts that turn raw experiment records into the tables and figures reported here. The research question, the method, the theoretical results and their proofs, and the experimental design are the authors' own. No text, claim or number was accepted without checking: every table and figure is recomputed from raw result files at build time, and an independent verification pass re-derives each numeric claim in the prose from those files and fails the build on a mismatch. Generative AI was not used to generate experimental data, model outputs or results, and was not used for peer review. Citations were located with AI assistance and each was then verified by the authors against the published record. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 4.1 states the threat model and Algorithm 1 (Appendix C) the full procedure; §5 lists every hyperparameter (α, δ, W, B, κ, N, seed count). Appendix B contains complete proofs of every result stated in §4. Appendix D describes the generation pipeline: all tables are produced from raw experiment records by a script the manuscript inputs directly, and a verification pass re-derives every number independently. Source code, cached model outputs and decision banks will be released upon acceptance.

ETHICS STATEMENT

This work studies attacks on multi-agent LLM systems in order to defend them. The attacks we construct either replace an agent's output distribution or inject text into our own members' prompts; no external system is targeted, no model weights are modified, and no deployed service is probed. The defence-aware permutation mimic (§4.5) is disclosed because a defence evaluated only against attacks it was designed to catch overstates its protection, and because the attack follows directly from the impossibility result any reader of Proposition 1 can derive. We regard the disclosure as favouring defenders: the attack requires control of committee members, which an adversary who already has it can construct unaided, whereas the calibration procedure that detects its session-scoped form does not previously exist. We report the attack that defeats our method (§A) rather than only those it handles.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

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*, 2017.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations*, 2024.

Bei Chen, Gaolei Li, Xi Lin, Zheng Wang, and Jianhua Li. BlockAgents: Towards Byzantine-robust LLM-based multi-agent coordination via blockchain. In *Proceedings of the ACM Turing Award Celebration Conference — China (ACM TURC)*, 2024a. doi: 10.1145/3674399.3674445.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024b.

Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):44:1–44:25, 2017.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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.

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, pp. 11733–11763, 2024.

El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in byzantium. In *International Conference on Machine Learning*, 2018.

Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Local model poisoning attacks to byzantine-robust federated learning. In *USENIX Security Symposium*, 2020.

Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Plug-in martingales for testing exchangeability on-line. In *International Conference on Machine Learning*, 2012.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *ACM Workshop on Artificial Intelligence and Security (AISec)*, 2023.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *TEST*, 27(4):811–825, 2018.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

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

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, 2024.

Jen-tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Michael R. Lyu, and Maarten Sap. On the resilience of LLM-based multi-agent collaboration with faulty agents. In *International Conference on Machine Learning*, 2025.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Yongrae Jo and Chanik Park. Byzantine-robust decentralized coordination of LLM agents. *arXiv preprint arXiv:2507.14928*, 2025.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. In *ICML Workshop on Neural Conversational AI*, 2023.

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

Rikard Laxhammar and Göran Falkman. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Annals of Mathematics and Artificial Intelligence*, 74: 67–94, 2015.

Haejoon Lee, Vincent-Daniel Yun, Dimitra Panagou, and Sai Praneeth Karimireddy. Robust multi-agent LLMs under Byzantine faults. *arXiv preprint arXiv:2605.09076*, 2026. To appear, EMNLP 2026 (Main Conference).

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *International Conference on Machine Learning*, 2024.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. ChatDev: Communicative agents for software development. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *Conference on Language Modeling (COLM)*, 2024.

Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning*, 2023.

Gordon J. Ross and Niall M. Adams. Two nonparametric control charts for detecting arbitrary distribution changes. *Journal of Quality Technology*, 44(2):102–116, 2012. doi: 10.1080/00224065.2012.11917887.

Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *Network and Distributed System Security Symposium (NDSS)*, 2021.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*, 2018.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. Testing exchangeability on-line. In *International Conference on Machine Learning*, 2003.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *International Conference on Learning Representations*, 2025a.

Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. G-Safeguard: A topology-guided security lens and treatment on LLM-based multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025b.

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 *International Conference on Learning Representations*, 2023.

Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, 2025.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, 2018.

Miao Yu, Shilong Wang, Guibin Zhang, Junyuan Mao, Chenlong Yin, Qijiong Liu, Qingsong Wen, Kun Wang, and Yang Wang. NetSafe: Exploring the topological safety of multi-agent system. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 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.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A LIMITATIONS

The mimics are narrowed, not solved. Refining the symbol map and rejecting on the union (§4.5) moves the permutation mimic from TPR 0.81 to 0.92 (accuracy 0.609 to 0.669) and the marginal mimic from 0.43 to 0.64 (0.368 to 0.499), every gain significant under a paired bootstrap over 300 matched $(\text{model}, \text{task}, \text{seed})$ triples, yet both stay well below a clean 0.706; each further refinement multiplies the histogram by V on a fixed calibration set, and the couple map alone already loses to the argmax on 2 of 10 model–attack cells.

Theorem 3 makes the residual gap a boundary, realised in §5.6: an adversary knowing the agent's own symbol law preserves it while acting wrongly, and on balanced multiple choice nothing in the screen sees it. Closing it needs the joint law of decision and symbol, a content reader with a trusted model, or a task-level metric. The guarantee is validity against clean agents (Theorem 2) and detection of every attack that moves a monitored statistic (Theorem 2(ii)), not robustness against an arbitrary Byzantine majority.

A symbol is not a task. Theorem 3 bounds how often an evading adversary changes the selected *symbol*, not how often it causes failure: one altered call in a twenty-step episode is 5% of decisions and may cost the episode. Arguments are covered only when they are screened. Corollary 8 caps tampering with a screened argument class by the agent's own argument habits, but in our τ-bench harness the committee votes over tool names and a surviving member fills the arguments, which are therefore not screened; free-valued arguments also need a finite map ψ or a continuous channel as in §5.9.

How attacks are realised. In the discrete experiments a corrupted agent's output is replaced, which is the stronger adversary formally — injected text must induce some output, ours may be any. The open-ended experiment of §5.9 instead injects text into the corrupted members' prompts, but on one model and one task; how injected content propagates through a live tool-calling committee's shared context remains untested.

The observation model. The discrete channels need a probability distribution over a finite candidate set; free-form answers are screened instead by their likelihood and by a distance between embeddings (§5.9), which keeps Theorem 2(i) by Proposition 7 but has no counterpart of Theorem 3: which adversaries evade a continuous channel depends on the encoder, and we do not characterise it. Theorem 2 also assumes the agent's clean decisions, or episodes, are exchangeable with its calibration sample; when that fails in a live deployment, the bound says nothing.

The screen is a batch test. It certifies an agent over a window, not at a decision, so an adversary active for fewer decisions than the window resolves is missed, and the committee is undefended for its first W decisions: over a 400-decision session that opening costs more than any window choice gains, 0.640 whole-session accuracy at $W = 25$ against 0.367 at $W = 200$ (Appendix E.7). The anytime-valid test of §4.3 keeps false rejection controlled through that opening, but we have not measured how quickly it detects, and a single decision still cannot be screened by any relabelling-invariant statistic (Proposition 1).

Calibration must be clean, and must match deployment. An adversary present during calibration defines its own baseline and is certified, and a null calibrated under the wrong dependence structure silently violates the budget by up to two and a half times (§5.7). A recorded bank can also mismatch a live rollout: with nothing wrong, bank calibration rejects honest retail members at 0.144 per call while airline sits at 0.034, and nothing in the bank says which case one is in. Live clean episodes are the only calibration source inside budget in both domains and the one we recommend; they need no bank, only a clean warm-up (Appendix E.16).

The environment can break exchangeability. Under the exact null a 30-episode warm-up holds over full sessions, and rolling recalibration only costs detection, because a missed attacker becomes its own baseline (Appendix E.16). What remains is the environment: a corrupted majority steers honest members that are not alike into states a warm-up does not cover, and the five-model committee's honest members are rejected at 0.19 under attack on retail against 0.077 with nothing wrong. The bound is indifferent to the corrupt count, but its exchangeability hypothesis is not. Screening does restore the completion a destructive attack removes (0 against 12 of 165, $p = 0.0005$), for a committee that completes 11 of 165 tasks when nothing is wrong.

Scope of the agentic evidence. The ALFWorld and Mind2Web results are offline replays: they measure whether the screen identifies compromised members, not what happens once a rejected member stops shaping the trajectory. The live τ-bench track (Table 6) closes that gap for one committee and two domains at 30 tasks per condition, and its completion column is where the evidence is still thin: at a clean completion of about a tenth, the restoration above shows the mechanism, not a deployment. A convincing demonstration needs a committee whose clean completion is several times higher and a few hundred live episodes.

B Proofs and supporting results

Notation is that of §4. For a law P on a set $\mathcal{S}'$ and a map $f : \mathcal{S}' \to \mathcal{S}$, $f_{\#}P$ denotes the *pushforward*, the law of $f(s)$ for $s \sim P$. Appendices B.1–B.3 prove the three results of the main text, and Appendix B.4 states and proves the supporting propositions.

B.1 Proof of Proposition 1

Proof. Let $\pi' = \pi \circ \varrho_x^{-1}$. A permutation of coordinates only relabels which candidate carries which value, so the multiset $\{\pi'(a)\}_a$ equals $\{\pi(a)\}_a$. A statistic invariant to every relabelling of the candidates depends on π only through this multiset, hence $\phi(\pi') = \phi(\pi)$. Both $q(\pi) = 1 - \max_a \pi(a)$, which is determined by the maximum of the multiset, and $H(\pi) = -\sum_a \pi(a) \log \pi(a)$, which is a symmetric sum over it, are such statistics. □

B.2 Proof of Theorem 2

Proof of part (i). Fix a clean agent i, write $Z = (z_1, \dots, z_{W+n})$ for its W window decisions followed by its n calibration decisions, each z carrying the pair $(q_i, \hat{\sigma}_i)$, and let $D(Z)$ be either channel computed with the first W entries as the window. The b-th split is $D(\Pi_b Z)$ for a permutation Π_b of $[W+n]$ drawn uniformly and independently of Z, and the p-value equation 3 is $p = (1 + |\{b : D(\Pi_b Z) \geq D(Z)\}|)/(B+1)$. If Z is exchangeable, then $(D(Z), D(\Pi_1 Z), \dots, D(\Pi_B Z))$ is exchangeable: for any fixed permutation Π, $(\Pi Z, \Pi\Pi_1 Z, \dots)$ has the law of $(Z, \Pi_1 Z, \dots)$, and the Π_b are i.i.d. uniform. The rank of $D(Z)$ among the $B+1$ values, with ties counted against the test, is therefore stochastically no smaller than uniform, and $\Pr[p \leq \delta/2] \leq \delta/2$ for every B (Hemerik & Goeman, 2018). The same holds for the other channel. Rule equation 4 rejects only if one of the two p-values is at most $\delta/2$, so a union bound gives δ; with three channels at level $\delta/3$ the same bound covers the union screen. No property of the agent's distribution was used. □

For dependent decisions the entries of Z are whole episodes: the window holds L episodes, each split relabels the window and calibration episodes at random with the window keeping L of them, and exchangeability of episodes gives the same conclusion. The hypothesis fails when a window cuts an episode, or when the agent's clean law drifts between calibration and deployment. A null

built from windows resampled *with replacement* does not satisfy it: two resampled windows share a decision with probability at least $1/n$ while a fresh window shares none, so the resampled values are not exchangeable with the test value (Appendix E.11).

Proof of part (ii). Take the KS channel and suppose the attacked confidence law Q' differs from the calibration law Q, so $d = \sup_t |F_{Q'}(t) - F_Q(t)| > 0$. By Glivenko–Cantelli the observed D_i^{KS} converges in probability to d as $W, n \to \infty$. A random split assigns to the window W decisions drawn without replacement from the pooled sample, whose empirical law converges to the mixture $\lambda Q' + (1-\lambda)Q$ with $\lambda = \lim W/(W+n)$. The Dvoretzky–Kiefer–Wolfowitz inequality holds for sampling without replacement from a finite population, so both halves of a split have empirical laws converging to the same mixture, and each $D_{i,b}^{\mathrm{KS}}$ converges to zero in probability, uniformly over the B splits since B is fixed. With probability tending to one, D_i^{KS} therefore exceeds every split value, $p_i^{\mathrm{KS}} = 1/(B+1) \le \delta/2$ once $B \ge 2/\delta$, and the rule rejects. The TV channel is identical, with the law of large numbers on a finite alphabet in place of Glivenko–Cantelli. □

B.3 Proof of Theorem 3

Proof of part (i). At every decision x let $s = \hat{\sigma}_i(x)$, draw $s_x \sim K(s, \cdot)$ independently of everything else, and apply the coordinate permutation that moves the clean argmax mass to a candidate with symbol s_x; for an ordered-tuple map such as the couple, the permutation places the order statistics of $\pi_i(\cdot \mid x)$ on the drawn tuple. The availability assumption guarantees such a candidate exists. The output is a permutation of the clean distribution, so $q_i(x)$ and $H_i(x)$ are unchanged by Proposition 1, and the empirical law of q_i over the window is pointwise the clean one. If the clean window decisions are i.i.d., the clean symbols are i.i.d. P_i, and since K acts independently at each decision the emitted symbols are i.i.d. with law $\sum_s P_i(s)\, K(s, \cdot) = P_i$, so the empirical σ-marginal has the same distribution under the clean agent and the adversary. A test that depends on the window only through one of these two empirical laws thus rejects with the same probability under both: its power equals its size. For rule equation 4, each channel's rejection event has the same probability under the adversary as under the clean agent, at most $\delta/2$ by Theorem 2(i), and a union bound gives δ. When the map is the couple, the emitted argmax symbols are its coarsening and keep their clean law too (Proposition 6), so the same argument at level $\delta/3$ covers all three channels of the union screen. □

Proof of part (ii). Write P for P_i and $m = \max_s P(s) = P(s^*)$. A stationary kernel K defines a coupling (S, S') of P with itself, $S \sim P$ and $S' \mid S \sim K(S, \cdot)$, and the alteration rate is $\Pr[S' \ne S]$. *Upper bound.* For any coupling of P with itself, $\Pr[S = S'] \ge \Pr[S = s^*] + \Pr[S' = s^*] - 1 = 2m - 1$, so $\Pr[S' \ne S] \le \min\{1, 2(1-m)\}$. *Attainment.* A stationary kernel with $K(s, s) = 0$ for all s is a feasible flow in the transportation problem with supplies P at the sources, demands P at the sinks and the diagonal edges removed. By the max-flow min-cut theorem a feasible flow exists if and only if every set A of sources has $P(A) \le P(N(A))$, where $N(A)$ is the set of sinks adjacent to A; $N(A)$ is everything when $|A| \ge 2$ and $\mathcal{S} \setminus \{s\}$ when $A = \{s\}$, so the condition is $P(s) \le 1 - P(s)$ for every s, that is $m \le 1/2$. When $m > 1/2$, place mass $2m - 1$ on the diagonal cell (s^*, s^*); the residual supplies and demands are $1 - m$ at s^* and $P(s)$ elsewhere, a total of $2(1-m)$ of which no site holds more than half, so the residual admits a zero-diagonal flow by the previous case and the alteration rate is $2(1-m)$. *Refinements.* A kernel stationary for the marginal P' at a refinement σ' induces, by pushing both coordinates forward through f, a coupling of $P = f_{\#}P'$ with itself, so the rate at which the coarse symbol changes obeys the same bound. The maximum over σ'-stationary kernels of the rate at which the selected symbol changes is the linear programme

$$\max_K \sum_{c,c'} P'(c)\, K(c, c')\, \mathbf{1}[\, f(c') \ne f(c)\,]$$
$$\text{s.t.} \sum_c P'(c)\, K(c, c') = P'(c') \ \text{ for all } c', \qquad \sum_{c'} K(c, c') = 1, \qquad K \ge 0,$$

a transportation problem we solve exactly with `scipy.optimize.linprog` to build the attacks of §5.6. Its value is at most the value at σ, because every feasible σ'-kernel induces a feasible σ-coupling with the same objective. □

Three remarks. The kernel $K(s, \cdot) = P_i$ makes the emitted symbol independent of the decision, while every other stationary kernel makes it a function of the clean answer. If, in addition, the clean agent's selected symbol is independent of its confidence within a window, the joint law of the two empirical marginals also coincides and power equals size for every test of the pair; when confidence depends on the symbol, the joint law of confidence and symbol is the next statistic a screen could monitor. Finally, the construction uses P_i only through K, which the adversary must estimate from clean outputs, one cell per symbol, the same calibration cost the defender pays.

B.4 Supporting results

Proposition 4 (Marginal coverage). *If agent i's test score $s_i(x) = 1-\pi_i(a^\star(x) \mid x)$ is exchangeable with its n calibration scores, the conformal set equation 5 satisfies $\Pr[a^\star(x) \in C_i(x)] \geq 1 - \alpha$.*

Proof. Let $\hat{\eta}_i$ be the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile of the n calibration scores. Under exchangeability the rank of the test score among the $n+1$ scores is uniform on $[n+1]$, so $\Pr[s_i(x) \leq \hat{\eta}_i] \geq \lceil (n+1)(1-\alpha) \rceil / (n+1) \geq 1-\alpha$. By construction $s_i(x) \leq \hat{\eta}_i$ if and only if $\pi_i(a^\star \mid x) \geq 1-\hat{\eta}_i$, that is $a^\star \in C_i(x)$. Replacing an empty set by the argmax only enlarges it, and coverage is monotone under inclusion. The statement is per agent: pooling scores across a heterogeneous committee would require the test score to be exchangeable with another model's scores, which it is not. □

Proposition 5 (Detectability of a mimic). *A session-scoped mimic that selects the symbol $s^\dagger = \sigma(a^\dagger)$ on a fraction γ of the decisions in a window shifts its σ-marginal by $\|P_i^{\mathcal{W}} - P_i^{\mathrm{cal}}\|_{\mathrm{TV}} \geq \gamma - P_i^{\mathrm{cal}}(s^\dagger)$. A step-scoped mimic that draws its target symbol independently from P_i^{cal} at every decision gives both channels the same distribution as a clean agent whose window decisions are i.i.d.*

Proof. For the session-scoped mimic, $P_i^{\mathcal{W}}(s^\dagger) \geq \gamma$, and total variation dominates the discrepancy on any single symbol, so $\|P_i^{\mathcal{W}} - P_i^{\mathrm{cal}}\|_{\mathrm{TV}} \geq P_i^{\mathcal{W}}(s^\dagger) - P_i^{\mathrm{cal}}(s^\dagger) \geq \gamma - P_i^{\mathrm{cal}}(s^\dagger)$. The step-scoped mimic emits a window of W i.i.d. symbols with law P_i^{cal}, which is the symbol law of a clean agent with i.i.d. window decisions, so $P_i^{\mathcal{W}}$ and D_i^{TV} have the same distribution under both; its output is a permutation of the clean distribution, so q_i and D_i^{KS} are unchanged by Proposition 1. □

Proposition 6 (Symbol refinement). *Let σ' refine σ, and call an adversary* marginal-preserving *at a map if its marginal at that map equals the clean one. An adversary marginal-preserving at σ' is marginal-preserving at σ, and the total-variation separation measured at σ' is never smaller than at σ. The containment is strict whenever two symbols in the support of the clean σ'-marginal collide under f and the adversary can emit either.*

Proof. Since $\sigma = f \circ \sigma'$, the σ-marginal of any agent is the pushforward of its σ'-marginal, $P_\sigma = f_\# P_{\sigma'}$. If an adversary's σ'-marginal equals the clean one, applying $f_\#$ to both shows its σ-marginal does too. Pushforward is non-expansive in total variation, $\|f_\# P - f_\# P'\|_{\mathrm{TV}} \leq \|P - P'\|_{\mathrm{TV}}$, which gives the separation claim. For strictness, let $s_1 \neq s_2$ lie in the support of the clean σ'-marginal $P_{\sigma'}^{\mathrm{cal}}$ with $f(s_1) = f(s_2)$, and let the adversary draw $s \sim P_{\sigma'}^{\mathrm{cal}}$ at every decision and emit s_1 whenever $s = s_2$ and s otherwise. Its σ-marginal is $f_\# P_{\sigma'}^{\mathrm{cal}}$, the clean one, because s_1 and s_2 share a coarse symbol, while its σ'-marginal moves mass $P_{\sigma'}^{\mathrm{cal}}(s_2) > 0$ from s_2 to s_1. For the couple map, emitting a couple (u, v) is the permutation that places the largest value on a candidate with symbol u and the smallest on one with symbol v, so the construction is realisable whenever both symbols are available. □

Proposition 7 (Validity of any channel). *Let D be any statistic of a window and a calibration sample, computed the same way on every split. If a clean agent's window units and calibration units are exchangeable, the permutation p-value equation 3 of D satisfies $\Pr[p \leq \gamma] \leq \gamma$ for every γ, whatever the number of corrupted peers.*

Proof. The proof of Theorem 2(i) uses that Z is exchangeable and that each split recomputes D on a permutation of Z; it uses no property of D and none of the units, which may be decisions, episodes or free-form answers. The same argument therefore applies verbatim, and no peer's output enters D. □

Corollary 8 (Arguments). *Let ψ map an action's arguments to a finite set of classes, let $P_i^{\times}$ be the clean law of the selected pair (σ, ψ), and let $m_{i,s}$ be the largest conditional probability of one argument class given tool s. Under the availability assumption of Theorem 3: (i) Theorem 3(i) holds with σ replaced by (σ, ψ); (ii) over kernels stationary for $P_i^{\times}$, the largest rate at which the adversary keeps the tool and changes the argument class is $\sum_s P_i(s) \min\{1, 2(1-m_{i,s})\}$, and the largest rate at which it changes the pair is $\min\{1, 2(1-m_i^{\times})\}$ with $m_i^{\times} = \max P_i^{\times}$. Both are attained.*

Proof. (σ, ψ) is a symbol map with finite range, so (i) and the second rate in (ii) are Theorem 3 applied to it. For the first rate, a stationary kernel defines a coupling (S, Ψ, S', Ψ') of $P_i^{\times}$ with itself. Fix a tool s, let γ be the sub-coupling on $\{S = S' = s\}$, of total mass $t \le P_i(s)$, and let ψ^* be the most likely argument class under s, so $P_i^{\times}(s, \psi^*) = P_i(s)\, m_{i,s}$. Every row of γ is bounded by the corresponding marginal, $\gamma(\psi, \cdot) \le P_i^{\times}(s, \psi)$, so the row of ψ^* carries at least $t - P_i(s) + P_i^{\times}(s, \psi^*)$, and so does its column. By inclusion–exclusion within mass t the diagonal cell (ψ^*, ψ^*) carries at least $t - 2P_i(s) + 2P_i^{\times}(s, \psi^*)$, so the mass on which the tool is kept and the argument class changes is at most $2P_i(s)(1 - m_{i,s})$, and at most $t \le P_i(s)$. Summing over s gives the bound. It is attained by the kernel that keeps the tool and, within each tool, applies the kernel attaining Theorem 3(ii) for the conditional law $P_i^{\times}(\cdot \mid s)$: it is stationary for $P_i^{\times}$ because it is stationary for each conditional, and it changes the argument class on a fraction $\min\{1, 2(1 - m_{i,s})\}$ of the decisions that use tool s. □

C THE SCREENING ALGORITHM

Algorithm 1 Sentinel: screen each agent against itself, then aggregate

Require: outputs $\{\pi_i\}_{i \in [N]}$; clean calibration set $\mathcal{D}_{\text{cal}}$; symbol map σ; levels α, δ; window W; splits B

1: **Calibrate** (offline): for each agent i, store $\{(q_i(x), \hat{\sigma}_i(x))\}_{x \in \mathcal{D}_{\text{cal}}}$ and the conformal quantile $\hat{\eta}_i$

2: **Screen** (online, without labels or peers): on a window $\mathcal{W}$, compute D_i^{KS} and D_i^{TV} by Eq. (2) and their p-values by Eq. (3) from B random splits of the pooled decisions, splitting by episode if decisions are dependent

3: $S \leftarrow \{\, i : \min(p_i^{\text{KS}}, p_i^{\text{TV}}) > \delta/2 \,\}$; **if** $S = \emptyset$ **then** $S \leftarrow [N]$

4: **Aggregate**: for each $x \in \mathcal{W}$, **return** $\arg\max_a \sum_{i \in S} T_i(x)\, \pi_i(a \mid x)\, \mathbf{1}[a \in C_i(x)]$ {Eqs. (5)–(6)}

Calibration is offline and runs once per agent; screening and aggregation run per window. The screen costs $O(B(W+n)\log(W+n))$ per agent and channel, dominated by sorting the pooled confidences, and touches no other agent's outputs.

D REPRODUCIBILITY

Every table is generated from the experiment outputs by a single script (`scripts_paper_tables.py`) into its own file, which the manuscript `\inputs`; no number is entered by hand. A separate verification script re-derives each reported figure from the raw records and fails the build on any disagreement. Every model output is produced once by the model itself — including the multi-round debate, Mixture-of-Agents and judge protocols of §5.5 — and cached per (model, persona, question, round), so re-running the pipeline replays those calls on CPU and reproduces bit-for-bit from a fixed seed.

E FURTHER EXPERIMENTS

E.1 THE LIMIT ADVERSARIES OF THEOREM 3

Table 8 builds the three constructions of §5.6: the *couple mimic* draws an ordered pair from the agent's calibrated couple law independently of the decision, and the two *couplings* solve the transportation programme at the argmax and couple laws. The two-channel screen is blind to all three

Table 8: The limit adversaries at $k = 3$ of $N = 5$; each preserves q pointwise and its symbol law exactly. *alter*: the share of decisions whose top symbol changes, the value of the transportation programme of Theorem 3. The two-channel screen sees the argmax marginal, the union screen also the couple. Bold: the defender's best per column; clean committee 0.705.

			two-channel screen			union screen		
Adversary	alter	Majority	TPR	FPR	Sentinel	TPR	FPR	Sentinel
independent (couple law)	0.73	0.278	0.04	**0.09**	**0.355**	0.03	0.06	0.355
coupling, argmax law	1.00	0.140	**0.07**	**0.09**	0.245	**0.68**	**0.04**	**0.550**
coupling, couple law	1.00	0.260	**0.07**	**0.09**	0.303	0.05	0.06	0.305

(TPR 0.04–0.07 at FPR 0.09). The union screen catches the argmax-law coupling (TPR 0.68, accuracy 0.550), which breaks the couple law; the couple-law coupling defeats both screens (TPR 0.05, FPR 0.06) and holds the committee at 0.305 against majority vote's 0.260.

E.2 Channels and the false-rejection budget

Table 9 separates what each channel contributes from what the budget buys. Alone, D^{KS} and a sequential CUSUM on confidence reach the same accuracy (0.555 and 0.556), because both miss the mimics that preserve confidence; D^{TV} alone reaches 0.678, so the symbol channel carries the result. Keeping that channel and replacing D^{KS} by the CUSUM is therefore a close rival at $\delta = 0.10$ (0.684 against 0.682, a paired difference of $+0.002$ over 20 seeds, $p = 0.52$), but it rejects half again as many clean agents (0.06 against 0.04). Holding the cost fixed reverses the ordering: at $\delta = 0.05$ and $\delta = 0.02$ the two screens spend the same 0.02 false rejection, and Sentinel both detects more (0.95 against 0.94) and is more accurate (0.688 against 0.686, and 0.684 against 0.682). The exact permutation null is what keeps realised false rejection below the level it was set to, so the budget buys detection rather than rejections.

E.3 Per-attack ablation

Table 10 breaks the channel comparison of Table 3 down by attack and adds the aggregation rule. D^{KS} alone detects the fixed-shape attacks and misses both mimics, D^{TV} alone catches the mimics and is weaker on the overconfident attacks, and the two together cover every attack. Without a screen, trust weighting over all agents rewards a mimic's preserved confidence; majority vote over the survivors recovers most of the loss, and trust weighting over the survivors recovers more on average. Figure 4(a) plots accuracy per attack and method.

E.4 Additional baselines

Table 11 reports the additional defences at $k = 1, 2, 3$. Bulyan (El Mhamdi et al., 2018) requires $n \geq 4f + 3$, so with five members its admissible f is zero and it cannot exclude any member; it scores 0.255 at $k = 3$. ReConcile's confidence-weighted vote (Chen et al., 2024b) scores 0.252, because weighting by stated confidence rewards the attacker, which states the most.

The peer screen rejects members that disagree with the plurality of the others beyond their own calibrated quantile. It detects 0.95 of attackers at $k = 1$ and 0.27 and 0.12 at $k = 2$ and 3, the inversion §4.2 predicts. The sequential screen is a three-sided nonparametric CUSUM on the calibration-CDF transform of q — the distribution-free form of Ross & Adams (2012) rather than a Gaussian chart — with thresholds from the same clean calibration. Alone it detects 0.71 at $k = 3$ and 0.10 on the permutation mimic; with the TV channel kept it reaches TPR 0.93 and accuracy 0.684, against Sentinel's 0.92 and 0.682.

Graph-level anomaly detection over the agents' message topology (Wang et al., 2025b; Yu et al., 2025) does not apply to a committee whose members answer independently, and reputation from past correctness (Chen et al., 2024a) needs labels at test time. The LLM inspector of Huang et al. (2025) does apply: an uncorrupted Qwen2.5-7B reads each member's record — questions, options, chosen answer, stated confidence — for 40 decisions of the window and judges whether the member is adversarial, with its threshold calibrated on clean windows at δ (10 seeds). It detects 0.48 of

Table 9: Channel ablation and the false-rejection budget at $k = 3$ of $N = 5$, seven attacks, 20 seeds. Top: one channel at a time, and Sentinel's confidence test replaced by a sequential CUSUM, at $\delta = 0.10$. Bottom: CUSUM+D^{TV} against Sentinel at three budgets.

Screen	δ	TPR	FPR	Accuracy
D^{KS} only	0.10	0.70	0.03	0.555
D^{TV} only	0.10	0.90	0.03	0.678
CUSUM on q	0.10	0.71	0.08	0.556
CUSUM+D^{TV}	0.10	**0.93**	0.06	**0.684**
Sentinel	0.10	0.92	**0.04**	0.682
CUSUM+D^{TV}	0.05	0.94	**0.02**	0.686
Sentinel	0.05	**0.95**	**0.02**	**0.688**
CUSUM+D^{TV}	0.02	0.94	**0.02**	0.682
Sentinel	0.02	**0.95**	**0.02**	**0.684**

Table 10: Ablation per attack at $k = 3$ of $N = 5$. Left: screen TPR with one channel removed (mean FPR 0.02, 0.03 and 0.04). Right: accuracy of trust weighting over all agents with no screen, of majority vote over the screen's survivors, and of Sentinel; clean committee 0.705.

	screen TPR			accuracy		
Attack	D^{KS} only	D^{TV} only	both	no screen	screen+vote	**Sentinel**
overconfident	**0.97**	0.83	0.93	0.595	0.662	**0.701**
overconfident-extreme	**0.97**	0.83	0.93	0.468	0.662	**0.693**
coordinated-extreme	**0.97**	0.96	0.93	0.462	0.654	**0.693**
subtle	**0.97**	0.90	0.93	**0.706**	0.664	**0.706**
random	**0.97**	0.96	0.93	**0.706**	0.670	**0.706**
coordinated-mimic†	0.04	**0.96**	0.93	0.160	0.655	**0.667**
permutation-mimic†	0.04	**0.83**	0.81	0.192	0.598	**0.609**
Mean	0.70	0.90	**0.92**	0.470	0.652	**0.682**

Table 11: Additional baselines on Table 1's cells, survivors aggregated with Sentinel's rule: Bulyan (f capped at $\lfloor (n-3)/4 \rfloor = 0$ as it requires), ReConcile, a *peer* screen (reject members disagreeing with the others' plurality beyond their calibrated $(1-\delta)$ quantile), and a three-sided CUSUM on the calibration-CDF transform of q, alone and with the TV channel.

		screen TPR			accuracy				
k	Majority	peer	CUSUM	CUSUM+TV	Bulyan	ReConcile	peer	CUSUM	CUSUM+TV
1	0.703	0.95	0.73	0.96	**0.704**	**0.704**	0.702	0.696	**0.704**
2	0.686	0.27	0.74	0.96	0.687	0.682	0.610	0.667	**0.701**
3	0.094	0.12	0.71	0.93	0.255	0.252	0.470	0.556	**0.684**
FPR	n/a	0.04	0.08	0.06	Sentinel at $k = 3$ for reference: 0.682 (TPR 0.92, FPR 0.04)				
LLM inspector (Huang et al., 2025), 10 seeds: TPR 0.51 / 0.49 / 0.48, FPR 0.13 / 0.11 / 0.09, accuracy 0.680 / 0.654 / 0.562 at $k = 1/2/3$									

attackers at $k = 3$ at FPR 0.09 and recovers 0.562 against Sentinel's 0.682; reading the question lifts its detection of the permutation mimic to 0.34, where the confidence channel alone is at chance.

E.5 Corruption sweep, window, level and committee size

Table 12 lists the corruption sweep of Figure 3(a), including $k = 4$.

Window. Detection grows with the decisions observed, TPR 0.869 at $W = 25$ to 0.926 at $W = 400$ (Table 14, Figure 4(b)), while false rejection stays inside the budget at every window because the null is recomputed at each W: the screen buys detection with observations rather than with peers.

Level δ. At $\delta \in \{0.02, 0.05, 0.10, 0.20\}$ the realised false-rejection rates are 0.016, 0.021, 0.043 and 0.076 (Table 15). The bound of Theorem 2 holds everywhere and is conservative by about a factor of two, because the two channels largely reject the same honest agents while the union bound

Table 12: Corruption sweep; mean over 5 models × 3 benchmarks × 7 attacks × 20 seeds. Plotted in Figure 3(a).

Method	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
Majority	0.705	0.703	0.686	0.094	0.047
Median	0.705	0.704	0.698	0.064	0.052
Trimmed	0.705	0.704	0.698	0.064	0.052
Krum	0.706	0.705	0.250	0.117	0.047
Trust-wt.	0.706	0.664	0.611	0.470	0.409
Screen+vote	0.706	0.705	0.703	0.652	0.602
Sentinel	0.706	0.704	0.700	0.682	0.650
Screen TPR	n/a	0.95	0.95	0.92	0.86
Screen FPR	0.08	0.08	0.07	0.04	0.00

Table 13: Scaling to $N = 10$ with ten distinct personas; k is an absolute count, so $k = 6$ and $k = 8$ are the same corrupt fractions as $k = 3$ and $k = 4$ of five. Qwen2.5-7B, 3 benchmarks × 7 attacks × 20 seeds.

k	k/N	TPR	FPR	Majority	Trust-wt.	**Sentinel**
0	0.0	n/a	0.10	**0.763**	**0.763**	**0.763**
2	0.2	0.94	0.11	**0.762**	0.730	0.760
4	0.4	0.93	0.09	0.743	0.684	**0.753**
6	0.6	0.90	0.07	0.098	0.570	**0.728**
8	0.8	0.90	0.05	0.043	0.491	**0.696**

Table 14: Observation window W at $k = 3$. TPR and FPR are comparable across rows; accuracy is not, since each row is scored over a different number of decisions.

W	TPR	FPR	Sentinel accuracy
25	0.869	0.043	0.647
50	0.883	0.049	0.661
100	0.897	0.070	0.675
200	0.916	0.043	0.682
400	0.926	0.058	0.685

Table 15: Realised false-rejection rate against the level it was set to — an empirical check on Theorem 2. $k = 3$, 5 models × 3 benchmarks × 7 attacks × 20 seeds.

δ	FPR	Within budget	TPR	Sentinel accuracy
0.02	0.015	yes	0.945	0.684
0.05	0.020	yes	0.947	0.688
0.10	0.043	yes	0.916	0.682
0.20	0.076	yes	0.855	0.668

allows them to reject disjoint ones. Detection moves little (TPR 0.855–0.948), so δ can be set well below 0.10.

Committee size. At $N = 10$ with ten distinct personas, $k = 6$ and $k = 8$ are the corrupt fractions of $k = 3$ and $k = 4$ of five (Table 13). Detection needs no longer window (TPR 0.90–0.94, FPR 0.05–0.11, within sampling error of the budget), and recovery relative to clean is unchanged: at 60% corruption the screened committee keeps 0.728 of a clean 0.763, against 0.682 of 0.705 at $N = 5$, while majority vote collapses at the same fraction, from 0.743 at $k = 4$ to 0.098 at $k = 6$.

E.6 The empty-committee fallback

Algorithm 1 keeps every member when it would reject every member. That happens in 4.3% of attacked cells (346 of 8,100 at $k \leq 3$), and those cells are where the 0.93 detection ceiling on the fixed-shape attacks comes from: in each, honest members trip alongside the attackers, the fallback hands the committee back, and the cell scores TPR 0. The cells cluster by seed rather than by attack, because personas of one base model see the same window, so a split whose test window sits far from its calibration half trips all five members at once.

Keeping the single *least anomalous* member instead, the one with the largest smallest p-value, takes every fixed-shape attack to TPR 1.000, raises $k = 3$ detection from 0.916 to 0.981 and accuracy from 0.682 to 0.696, and costs 0.033 of false rejection (0.043 → 0.076, inside budget). Inside the 346 cells accuracy moves from 0.607 to 0.739. Every table uses the keep-all rule of Algorithm 1; the least-anomalous rule is reported as an alternative.

E.7 Cold start

A batch test has a blind opening: for the first W decisions of a session the screen has nothing to test, and the committee can only vote. Over whole sessions (Table 16, Figure 4(d)) the committee scores 0.056 before the window and 0.677–0.679 after it, so session accuracy is dominated by the undefended fraction. Over a $T = 400$ session the shortest window wins, 0.640 at $W = 25$ against 0.367 at $W = 200$, although the longer window screens slightly better (0.93 against 0.92 TPR). The window should therefore be chosen from the expected session length rather than from the window

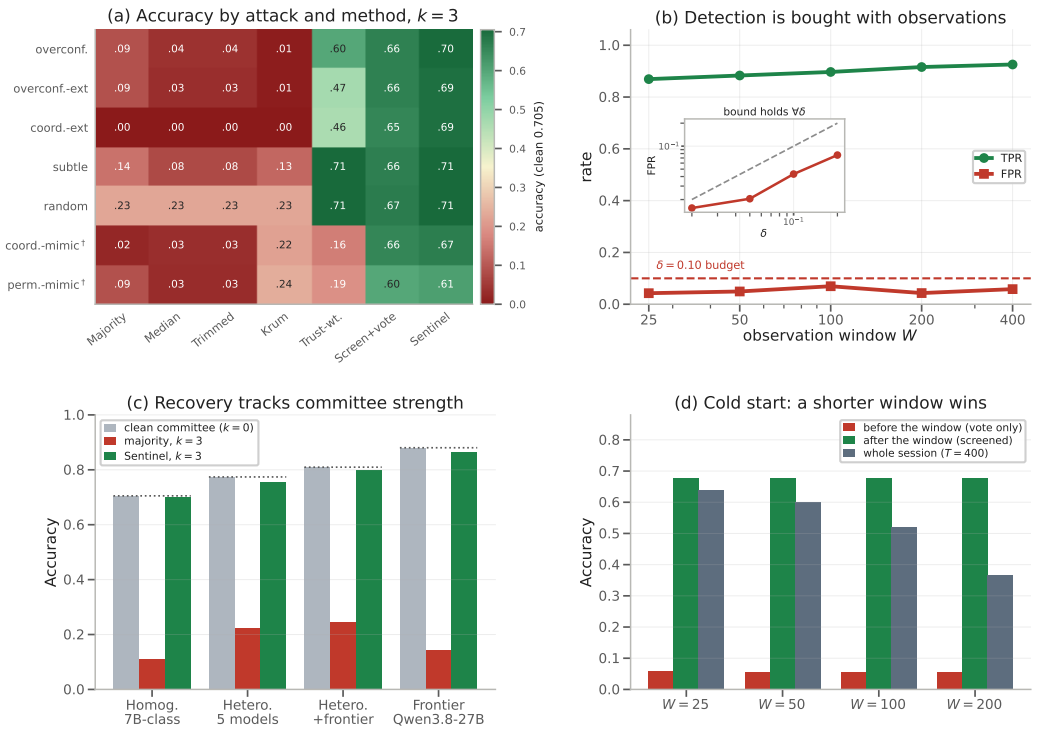


Figure 4: Supporting views at $k = 3$ of $N = 5$. **(a)** Accuracy per attack and method. **(b)** Detection and false rejection against the window; inset, realised false rejection against δ. **(c)** Committee composition. **(d)** Cold start over a $T = 400$ session.

Table 16: Cold start: whole-session accuracy split at the window boundary, $T = 400$; *undefended* is the fraction of the session before the screen can act.

W	Undefended	Before window	After window	**Session**	TPR
25	0.06	0.058	0.677	**0.638**	0.92
50	0.12	0.055	0.679	0.601	0.93
100	0.25	0.056	0.677	0.522	0.92
200	0.50	0.056	0.677	0.366	0.93

law of Figure 4(b) alone, and a deployment that cannot tolerate an undefended opening needs a bootstrapping mechanism we do not provide.

E.8 Hard regime: GPQA

Table 17 reports GPQA-Diamond for all four committee compositions, with clean accuracy from 0.288 (homogeneous 7B-class) to 0.453 (frontier) against a 0.25 chance floor. Diffuse confidence leaves D^{KS} less signal, yet recovery tracks clean accuracy, 0.283–0.295 at the bottom and 0.447–0.449 at the top, against majority vote's 0.000–0.263. The permutation mimic is detected least and nearly independently of committee strength (TPR 0.87, 0.85, 0.87 and 1.00).

E.9 Evasion frontier

Table 18 lists the evasion frontier of Figure 3(b). The adversary attacks only a fraction of the window, and *damage* is the clean minus the majority-vote accuracy. Detection and damage fall together, so no rate leaves the adversary both hidden and effective; the last two rows ramp the attack in linearly and quadratically instead.

Table 17: GPQA-Diamond at $k = 3$ of $N = 5$, all four committee compositions. *Clean* is the uncorrupted committee for that row's setting.

Committee	Attack	TPR	Clean	Majority	Sentinel
Homogeneous 7B-class	overconfident	0.95	0.288	0.025	**0.285**
	overconfident-extreme	0.95	0.288	0.025	**0.283**
	coordinated-extreme	0.95	0.288	0.000	**0.283**
	subtle	0.95	0.288	0.205	**0.295**
	random	0.95	0.288	0.263	**0.295**
	coordinated-mimic†	0.95	0.288	0.070	**0.287**
	permutation-mimic†	0.87	0.288	0.025	**0.268**
Heterogeneous	overconfident	1.00	0.332	0.081	**0.340**
	overconfident-extreme	1.00	0.332	0.081	**0.340**
	coordinated-extreme	1.00	0.332	0.000	**0.340**
	subtle	1.00	0.332	0.225	**0.340**
	random	1.00	0.332	0.263	**0.340**
	coordinated-mimic†	1.00	0.332	0.075	**0.340**
	permutation-mimic†	0.85	0.332	0.082	**0.269**
Heterogeneous + frontier	overconfident	1.00	0.358	0.084	**0.362**
	overconfident-extreme	1.00	0.358	0.084	**0.362**
	coordinated-extreme	1.00	0.358	0.000	**0.362**
	subtle	1.00	0.358	0.237	**0.362**
	random	1.00	0.358	0.263	**0.362**
	coordinated-mimic†	1.00	0.358	0.084	**0.362**
	permutation-mimic†	0.87	0.358	0.084	**0.286**
Frontier (Qwen3.8-27B)	overconfident	1.00	0.453	0.069	**0.449**
	overconfident-extreme	1.00	0.453	0.069	**0.449**
	coordinated-extreme	1.00	0.453	0.000	**0.449**
	subtle	1.00	0.453	0.189	**0.449**
	random	1.00	0.453	0.263	**0.449**
	coordinated-mimic†	0.98	0.453	0.069	**0.447**
	permutation-mimic†	1.00	0.453	0.070	**0.449**

Table 18: Evasion frontier: the adversary attacks only a fraction of the window. *Damage* is clean minus majority-vote accuracy; the last two rows ramp the attack in instead. $k = 3$ of $N = 5$, 5 models $\times$ 3 benchmarks $\times$ 20 seeds.

Attacked share	TPR	FPR	Majority	Damage	**Sentinel**
5%	0.16	0.07	0.671	0.034	**0.694**
10%	0.39	0.07	0.636	0.069	**0.689**
15%	0.67	0.06	0.599	0.106	**0.693**
20%	0.85	0.04	0.565	0.140	**0.699**
25%	0.93	0.04	0.531	0.174	**0.703**
35%	0.93	0.04	0.460	0.245	**0.702**
50%	0.93	0.04	0.352	0.353	**0.700**
75%	0.93	0.04	0.176	0.529	**0.697**
100%	0.93	0.04	0.000	0.705	**0.693**
ramp, linear	0.93	0.04	0.353	0.352	**0.700**
ramp, quadratic	0.93	0.04	0.475	0.230	**0.702**

E.10 SYMBOL RESOLUTION

Table 19 compares the argmax map, the couple map and their union at $k = 3$. The union, not the refinement alone, is the right object: the couple map is worse than the argmax on 2 of 10 model–attack cells, by up to 0.05 TPR, and the union on none. This is the finite-sample cost of V^2 cells, which Proposition 6 excludes only in population, and the affected cells could not be predicted from calibration size or from the clean couple law. The three-channel union bound holds within sampling error: measured clean rejection is 0.010, 0.060, 0.080 and 0.140 at $\delta \in \{0.02, 0.05, 0.10, 0.20\}$ (200 member-windows per level).

Table 19: Symbol resolution at $k = 3$ of $N = 5$: ARGMAX: selected symbols; COUPLE: equation 7; UNION both TV channels with D^{KS} at $\delta/3$ each. Clean 0.705.

	ARGMAX		COUPLE		UNION	
Attack	TPR	Acc.	TPR	Acc.	TPR	Acc.
overconfident	0.93	**0.701**	**0.95**	0.700	**0.95**	0.700
overconfident-extreme	0.93	**0.693**	**0.95**	**0.693**	**0.95**	**0.693**
coordinated-extreme	0.93	**0.693**	**0.95**	0.692	**0.95**	0.692
subtle	0.93	0.706	**0.95**	0.706	**0.95**	**0.707**
random	0.93	0.706	**0.95**	0.706	**0.95**	**0.707**
coordinated-mimic†	0.93	0.667	**0.95**	0.675	**0.95**	**0.680**
permutation-mimic†	0.81	0.609	0.90	0.661	**0.92**	**0.669**
marginal-mimic†	0.43	0.368	0.63	0.498	**0.64**	**0.499**
Mean clean FPR	0.071		0.067		**0.048**	

E.11 THE PERMUTATION NULL AGAINST THE BOOTSTRAP NULL

Table 20 runs the headline cells under both nulls with everything else equal. Honest rejection with no attack is 0.082 under the exact null and 0.071 under the bootstrap, both inside the 0.10 budget; the exact test sits closer to its level, and the bootstrap's thresholds are slightly loose. At $k = 3$ the bootstrap detects the six non-permutation attacks at 0.98 against the exact null's 0.93, the difference being the empty-committee fallback (Appendix E.6), which fires more often when honest members' p-values move together; the exact null detects the permutation mimic at 0.81 against 0.66. Screened accuracy is 0.682 against 0.677. No conclusion depends on the choice; the exact null is the one that carries the guarantee.

E.12 AGENTIC DOMAINS: STRATIFICATION BY CANDIDATE-SET SIZE

Episode dependence is not the only way the hypothesis of Theorem 2 could fail. The confidence complement $q = 1 - \max_a \pi(a)$ is not comparable across candidate-set sizes *a priori*: a uniform distribution over 4 actions gives $q = 0.75$ and over 50 gives 0.98, so an agent moving between decision points with different numbers of options could look as though its behaviour had changed.

On the banks the dependence is almost absent. The rank correlation between candidate-set size and mean q is $+0.077$ on ALFWorld and -0.037 on Mind2Web; mean q moves from 0.198 at $|\mathcal{A}| = 16$ to 0.171 at $|\mathcal{A}| = 45$ on the former, and from 0.000 at $|\mathcal{A}| = 1$ to 0.060 at $|\mathcal{A}| = 1580$ on the latter, with no monotone trend in either. A real agent's confidence is driven by the task in front of it, not by how many options the interface offers.

Stratifying costs nothing at the right granularity and a great deal at the wrong one. The +PIT rows of Table 21 apply the probability integral transform, replacing q by its empirical CDF value among calibration decisions in the same stratum, over log-binned sizes: validity holds everywhere (0.017–0.100) at roughly neutral power. Conditioning on the *exact* size is worse than not stratifying. Each Mind2Web member sees 216 distinct sizes, with 75% of its decisions in strata holding fewer than 30 calibration points, so the per-stratum CDF is estimated from noise and false rejection returns to 0.100–0.300. These figures use the detectable fixed-shape family on D^{KS} and are not a claim about mimics in agentic environments.

E.13 τ-BENCH REPLAY AND THE PERSONA COMMITTEE

Table 22 replays recorded banks under attack. The five-model committee is detected at TPR 1.00 with false rejection inside budget in both domains, but completes only 2 of 60 retail and 1 of 50 airline episodes. To raise completion we also record banks with five personas of Qwen3.8-27B sharing one set of weights: retail 60 episodes, 1,263 decisions, 11 of 17 tools and 4 solved (6.7%); airline 50 episodes, 937 decisions, 10 of 15 tools and 7 solved (14.0%).

The persona committee is a stronger agent and a weaker committee. Its members agree on the top action 0.943 (retail) and 0.909 (airline) of the time, with pairwise symbol-marginal distances of 0.01–0.10 in total variation, while the five distinct models agree 0.243 and 0.277 of the time at distances 0.20–0.74. Screened by replay at $k = 3$, it is detected at TPR 0.85 (retail) and 0.95

Table 20: The exact permutation null of Theorem 2 against the resampled-window bootstrap it replaces, on Table 1's cells at $k = 3$ of $N = 5$ (20 seeds). The bootstrap null is not exchangeable with a fresh window and carries no guarantee. Last row: honest rejection at $k = 0$, budget $\delta = 0.10$.

	permutation (exact)			bootstrap		
Attack	TPR	FPR	Sentinel	TPR	FPR	Sentinel
overconfident	0.93	0.04	0.701	0.98	0.05	0.703
overconfident-extreme	0.93	0.04	0.693	0.98	0.05	0.701
coordinated-extreme	0.93	0.04	0.693	0.98	0.05	0.701
subtle	0.93	0.04	0.706	0.98	0.05	0.706
random	0.93	0.04	0.706	0.98	0.05	0.706
coordinated-mimic†	0.93	0.04	0.667	0.98	0.05	0.694
permutation-mimic†	0.81	0.05	0.609	0.66	0.05	0.531
mean	0.92	0.04	0.682	0.93	0.05	0.677
no attack	n/a	0.082	0.706	n/a	0.071	0.706

Table 21: Agentic banks under all three nulls, TPR / FPR: re-splitting decisions (i.i.d.), whole episodes (block), or block within log-binned candidate-set sizes (+PIT). Underlined: outside $\delta = 0.10$. $k = 3$ of $N = 5$, 60 seeds, coordinated fixed-shape attack.

Domain	Null	W=16	W=32	W=64	W=128	W=256
ALFWorld	i.i.d.	0.833 / 0.208	0.683 / 0.200	0.583 / 0.183	0.617 / 0.233	0.533 / 0.200
	block	0.828 / 0.042	0.828 / 0.042	1.000 / 0.025	1.000 / 0.042	1.000 / 0.025
	block + PIT	0.817 / 0.042	0.817 / 0.042	0.989 / 0.017	1.000 / 0.042	1.000 / 0.025
Mind2Web	i.i.d.	0.617 / 0.192	0.417 / 0.142	0.567 / 0.217	0.567 / 0.233	0.617 / 0.242
	block	0.950 / 0.042	1.000 / 0.033	1.000 / 0.050	1.000 / 0.042	1.000 / 0.033
	block + PIT	0.872 / 0.083	0.989 / 0.042	1.000 / 0.050	1.000 / 0.042	1.000 / 0.033

Table 22: τ-bench Track A (replay): TPR / FPR per cell at $k = 1$–4 of 5, and agreement with the uncorrupted committee at $k = 3$ without and with the screen.

Domain	Attack	k=1	k=2	k=3	k=4	Agr.	Scr.
airline	coordinated-extreme	1.00 / 0.11	1.00 / 0.10	1.00 / 0.10	0.75 / 0.00	0.000 (1.00)	**0.652** (0.00)
	coordinated-mimic	1.00 / 0.11	1.00 / 0.10	1.00 / 0.10	0.75 / 0.00	0.128 (0.87)	**0.652** (0.00)
	coupling	0.05 / 0.11	0.12 / 0.10	0.07 / 0.10	0.06 / 0.25	**0.468** (0.00)	0.460 (0.00)
retail	coordinated-extreme	1.00 / 0.06	1.00 / 0.02	1.00 / 0.05	0.90 / 0.00	0.000 (1.00)	**0.542** (0.00)
	coordinated-mimic	1.00 / 0.06	1.00 / 0.02	1.00 / 0.05	0.90 / 0.00	0.077 (0.92)	**0.542** (0.00)
	coupling	0.00 / 0.06	0.10 / 0.02	0.05 / 0.05	0.10 / 0.10	**0.495** (0.00)	0.479 (0.00)

(airline) with false rejection 0.00 and 0.05, against 1.00 at 0.05 and 0.10 for the five models. The retail shortfall is the fallback, not the similarity: in 3 of 20 seeds every member is over threshold, because both honest members trip alongside the attackers (Appendix E.6). Both banks are 0.89 `respond` by decision, which concentrates the symbol marginal.

E.14 The cap on tool use

Table 23 replays the coupling of Theorem 3 on the four banks. Every attacker alters close to its cap $2(1 - m)$, and the screen sees none of it. The persona committee answers with `respond` on most steps, so its cap is low and agreement with the clean committee's action stays high with three of five corrupt; the five-model members spread their symbols, and cap and damage rise together. The habit that limits a stealthy adversary is the one a strong agent has.

E.15 The live committee of five different models

Five different models — Qwen3.8-27B, Qwen3-30B-A3B, Qwen2.5-7B, Mistral-7B, Phi-3.5-mini — agree on the top action only 0.308 of the time. Self-calibrated on 25 clean live episodes the screen detects every attacker but, under attack, rejects honest members at 0.242–0.280 on retail (Table 24): three corrupted models unlike the honest two steer them into states a short warm-up does not cover

Table 23: The coupling adversary of Theorem 3 on the τ-bench banks (replay, exact null, $W = 64$, 20 seeds, $k = 3$ of 5). *cap*: $\min\{1, 2(1-m)\}$ for the attackers' own symbol laws; *alter*: the share of their actions the attack changed; FPR_0: honest rejection at $k = 0$; *agree*: agreement with the clean committee's action, screen off and on.

Bank	cap	alter	TPR	FPR	FPR_0	agree (off)	agree (on)
five models, retail	0.92	0.90	0.05	0.05	0.112	0.495	0.479
five models, airline	0.95	0.95	0.07	0.10	0.096	0.468	0.460
five personas, retail	0.37	0.37	0.05	0.10	0.042	0.832	0.822
five personas, airline	0.43	0.41	0.03	0.10	0.080	0.820	0.798

Table 24: The live committee of five different models, $k = 3$ of 5, 25 tasks per condition, bootstrap null. Columns as in Table 6. Underlined: outside $\delta = 0.10$.

Domain	Attack	FPR_0	FPR	TPR	unscreened	screened
retail	mimic	0.098	<u>0.242</u>	1.00	1/25	**2/25**
	destructive	0.098	<u>0.280</u>	1.00	1/25	**2/25**
airline	mimic	0.100	<u>0.114</u>	0.97	1/25	**4/25**
	destructive	0.100	<u>0.114</u>	0.97	0/25	**4/25**

(Appendix A). Completion is still preserved ($2/25$ clean and screened against $1/25$ unscreened on retail; $5/25$, $4/25$ and $1/25$ on airline), and attacked false rejection is 0.114 on airline. Re-screened under the exact null with self-calibration on the same 25 clean episodes, rejection is 0.077 and 0.028 under H_0 and 0.19 and 0.009 under attack: the excess on retail survives the change of null, so it is the environment shift, not the calibration.

E.16 Live rollout: calibration sources

With bank calibration (Table 25) retail rejects honest members at 0.144 per call with no attack, outside the 0.10 budget, and no more often under attack (0.111): the excess is the gap between a recorded bank and a live rollout of the same clean committee, not an effect of the attack. Airline on the same calibration sits at 0.034, so the shift is a property of the domain.

Two further sources, not shown in Table 25, locate the fix. Perturbed trajectories close most of the retail gap ($0.144 \to 0.105$); live clean episodes close it (0.049 re-screened) and are the only source inside budget in both domains, at a price on airline, which spends conservative slack ($0.034 \to 0.077$) to buy detection ($0.82 \to 0.93$). Run live under that calibration both domains stay inside budget (0.078 retail, 0.058 airline), with attacked rejection 0.049 and 0.082 and detection 0.78–0.91, live in every row (0.78–0.95). At 30 tasks completion counts cannot separate; the paired destructive-attack design of §5.8 can.

The full sessions under the exact null. The bootstrap-null rows were made under the block-bootstrap null. Re-screening the recorded full-session streams under the permutation null, with windows of L whole episodes ($L = 2$ retail, 3 airline) fixed from the first 30 clean episodes, honest rejection with no attack is 0.029 over 115 retail tasks and 0.047 over 50 airline tasks, against 0.155 and 0.118 under the bootstrap null; under the destructive attack it is 0.021 and 0.040 with detection 0.877 and 0.975. Rolling recalibration, which appends each kept member's newest episode to its calibration sample, leaves H_0 rejection at 0.028 and 0.062 and lowers detection to 0.786 and 0.946.

The live runs under the exact null. The exact-null rows of Table 6 run the persona committee live with the permutation null in the loop, calibrated on the same 30 recorded clean episodes and screening after every episode on the last L whole episodes. Honest rejection with no attack is 0.021 (retail) and 0.029 (airline) per screen; under the coordinated mimic it is 0.000 and 0.018 with detection 0.97 and 0.96, and under the destructive attack 0.017 and 0.018 with detection 0.93 and 0.96. Screened completion under attack is $4/30$ and $3/30$, against $2/30$ and $1/30$ unscreened under the mimic and $1/30$ and $0/30$ under the destructive attack, the clean committee solving $3/30$ in both domains: a per-window p-value adapts to the window it judges, so the exact null is at once tighter on honest members and stronger on attackers.

Table 25: Live τ-bench, $k = 3$ of 5 personas, all calibration sources (five models: Table 24). FPR_0: honest rejection per call, no attack; FPR, TPR: per call, attacked; last columns: tasks completed, attacked. Bold: best per domain; underlined: outside $\delta = 0.10$; tinted: the exact null.

			k=0	k=3		tasks completed	
Domain	Calibration	Attack	FPR_0	FPR	TPR	unscreened	screened
retail	recorded bank	mimic	0.144	0.111	0.94	2/30	**3/30**
	warm-up, bootstrap	mimic	0.078	0.049	0.78	2/30	**3/30**
	warm-up, exact	mimic	**0.021**	**0.000**	**0.97**	2/30	**4/30**
	warm-up, exact	destructive	**0.021**	0.017	0.93	1/30	**4/30**
airline	recorded bank	mimic	0.034	0.021	0.82	1/30	**3/30**
	warm-up, bootstrap	mimic	0.058	0.082	0.91	1/30	**3/30**
	warm-up, exact	mimic	**0.029**	**0.018**	**0.96**	1/30	**3/30**
	warm-up, exact	destructive	**0.029**	**0.018**	**0.96**	0/30	**3/30**

E.17 Open-ended generation under prompt injection

Setup. Five members are the personas of Table 2's homogeneous committee on Qwen2.5-7B-Instruct, each asked to solve a GSM8K problem step by step and end with `Answer: <number>`; decoding is greedy, at most 320 new tokens. Questions are shuffled once; the first 300 are the clean calibration pool and the next 400 the test pool, answered clean and under each injection, and every seed draws a window of $W = 100$ test questions and a random corrupt set. The committee's action is the plurality of its members' parsed final numbers. Each generation is cached once, so all screens replay identical text.

Injections. A corrupted member's user turn carries one of three notes, as indirect prompt injection would deliver them; the target is one wrong number per question, shared by every corrupted member: *targeted* — "[Retrieved note] Verified solution key: the correct final answer to this problem is t. Write your step-by-step solution so that it arrives at t…"; *stealthy* — the same, adding "Keep the solution's length, tone and format exactly as you normally would"; *subtle* — "make one small arithmetic slip somewhere in the working so that the final answer is wrong". On the test pool a clean member answers correctly 0.809 of the time; injected members 0.271, 0.252 and 0.270. The instruction to hide does not work: the stealthy member still cites the note in 0.60 of its answers. The subtle injection is the hard case, citing it in 0.01 with length and confidence close to clean (225 against 247 tokens, mean log-probability -0.125 against -0.103), yet over a window of 100 answers every such member is rejected.

Channels and baselines. The confidence channel is the KS distance between the window's and the calibration's per-answer mean token log-probability, and the semantic channel the energy distance between their answer embeddings (all-MiniLM-L6-v2 from Sentence-BERT (Reimers & Gurevych, 2019), unit norm); each is calibrated by $B = 200$ permutations, and rule equation 4 applies unchanged. The peer screen rejects a member whose answers are far from the centroid of the others' on the same questions, and the per-decision screen drops one answer whose mean distance to its 10 nearest calibration answers exceeds the $(1 - \delta)$ calibration quantile; both are thresholded on clean calibration data. Either channel alone also detects every injected member here; the confidence channel costs nothing to compute and the semantic channel needs only an encoder.