FROM EVIDENCE TO ACTION: MAS PREDICTION AND DECISION FOR LOW-PROBABILITY EVENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting low-probability events requires both preserving the order of magnitude of probabilities and identifying evidence dependencies and executing actions aligned with the specified loss objective. Existing evaluations of Multi-Agent System (MAS) prediction and collaboration rarely distinguish communication gains, information propagation, and action execution within a single task, and building agents from the same foundation model further limits what can be inferred about cross-model collaboration. We propose a paired framework benchmarked against computable conditional probabilities. In a virtual public-service scenario, we manipulate source relationships and the disclosure of private records and structural information, and use a new decision branch to separate probability substitution from action execution. We use four models to form Homogeneous MAS, which jointly constitute a Hybrid MAS where each virtual world uses four balanced compositions and two decoding runs, and all experiments are repeated on the same worlds. In the main experiment over 128 worlds, ordinary communication relative to independent ensembling yields a mean absolute $\log_{10}$ error difference of -0.0152 in the Homogeneous MAS and -0.01402 in the Hybrid MAS. After two rounds of communication, the availability rate of valid private raw records for ordinary communication and source verification in the Hybrid MAS approaches 84%. Through paired replication across compositions, verifiable information channels, and public loss objectives, this study shows that group consensus, probability accuracy, and action optimality require separate evaluation benchmarks.

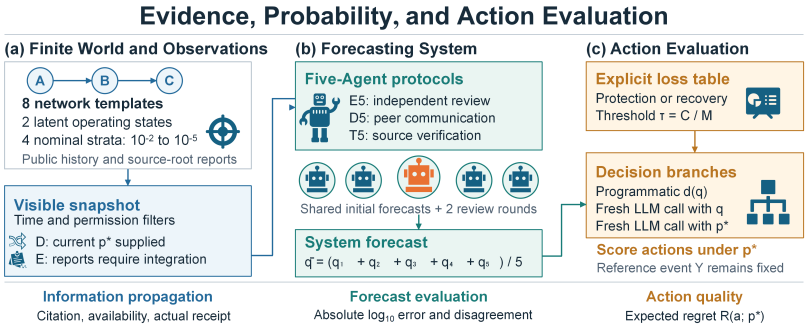

Figure 1: The finite worlds provide an evaluation reference only under identified information, evaluating MAS on low-probability forecasting and decision-making, which finds that while communication improves prediction accuracy, it doesn't ensure optimal actions.

1 INTRODUCTION

The difficulty for public-service systems is not identifying high-frequency events, but handling failures and their recovery mechanisms that occur with extremely low probability yet carry severe consequences. Large Language Models (LLM) have already been applied to real-event prediction and improve performance through retrieval augmentation, forecast aggregation, and task adaptation (Halawi et al., 2024; Karger et al., 2025). When the forecasting task is delegated to a LLM or a MAS,

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.

however, the core questions become whether forecasts preserve the accuracy of probability orders of magnitude(Kim et al., 2026), whether multiple agents can correct one another, and what ultimately justifies the resulting action(Schoenegger & Park, 2023; Tang et al., 2026).

Evaluating low-probability events makes distinguishing these factors all the more necessary: if the true probability is 10^{-5} and the forecast is 10^{-3}, the absolute error is below 0.001, yet the risk has been overestimated by two orders of magnitude. Black-Swan events also involve generating mechanisms that are not contained in the current information model:Tang et al. (2026) explain risk attribution for realized extreme trajectories along time, agent, and behavior dimensions. If two mechanisms present identical materials yet correspond to different true probabilities, adding agents with the same information cannot recover the missing structural information from nothing(Weng et al., 2025). Only once this boundary has been defined can forecast changes before and after disclosure be interpreted(Dorman et al., 2026; Ko et al., 2026).

We build an evaluation framework that runs through probabilities, evidence, and actions. First, finite Bayesian networks define the reference events, and conditional probabilities are computed from public history and source-identified reports, so that different architectures are compared within the same world and the same time point. Second, source independence and private records are manipulated in pairs, separately measuring system error under complete information, propagation of raw records, and recipients' deviation from a local-record reference. Third, with the decision task and the loss table held fixed, fresh decision calls are supplied with the old probability and with the correct probability respectively, testing the loss produced by forecast deviation and by execution of the specified objective. The structural-change experiment instead varies the disclosure time of a configuration certificate, linking identifiability analysis to forecasting across time points.

Agent composition is another necessary condition for testing communication gains, since agents produced by the same foundation model may share computational habits and error patterns. Several models each forming a system and several models discussing together are different objects of collaboration. We therefore construct balanced Hybrid MAS: each of the four models occupies one seat, and one extra chairman seat is rotated in turn, forming four five-agent compositions. The independent-ensembling control in the Hybrid batch already aggregates heterogeneous probabilities, so the communication group's difference from it corresponds to the protocol change produced by adding peer information within the same composition.

Across 128 independent worlds, both the Homogeneous's and the Hybrid's average show improved error under ordinary communication, with D5$-$E5 of -0.015213 and -0.0140173, respectively. The availability rate of valid private raw records after two rounds is about 84%, allowing record arrival and output deviation from the local reference to be evaluated separately. When the same probability is supplied again, the language-model interface increases mean expected loss in every case. The Hybrid normalized execution difference is $+4.78747 \times 10^{-6}$, and the correct-probability replacement difference is -7.43421×10^{-6}. The two batches of action experiments also show that even when the same probability is supplied again, the language-model interface increases mean expected loss.

Our main contributions are as follows:

- We propose a low-probability MAS evaluation framework that runs through evidence, probabilities, and actions, using computable conditional probabilities and a public loss function to measure information propagation, forecast error, and action loss separately.
- We reveal the average forecasting gain of ordinary communication in both Homogeneous and Hybrid MAS, and find that error can still rise while group disagreement contracts, clarifying the separation between consensus and accuracy.
- Through a separate decision branch, we find that given the same probability, language-model decisions incur additional mean expected loss relative to programmatic optimization, revealing the difference between forecast accuracy and action optimality.

2 Related Work

Real-event forecasting research already covers tournament comparisons, retrieval augmentation, and dynamic evaluation (Halawi et al., 2024; Karger et al., 2025), and shows that information leakage

and resolution timing change measured performance (Paleka et al., 2026; Kapoor & Narayanan, 2023). Confidence in answer correctness (Kadavath et al., 2022; Tian et al., 2023), logical consistency of forecasts (Paleka et al., 2024), and event-probability accuracy have different targets. Cognitive and probabilistic-choice experiments identify measurable behavioral biases (Binz & Schulz, 2023; Macmillan-Scott & Musolesi, 2024; Hintze et al., 2025; Hertwig et al., 2004), but do not directly identify internal computations or a stable utility function. Work on strictly proper scoring rules and tail evaluation (Gneiting & Raftery, 2007; Lerch et al., 2017; Allen et al., 2023) supports retaining the complete forecast target. We measure numerical deviation against exact conditional probabilities and avoid selecting worlds by rare realized outcomes.

Independent aggregation (Schoenegger et al., 2024) and interactive discussion (Du et al., 2023; Chen et al., 2024) can improve particular tasks, but a strong single agent, voting, self-correction, and heterogeneous models are necessary controls (Zhang et al., 2025; Choi et al., 2026; Bertalanič & Fortuna, 2026). Kim et al. (2026) show that collaboration gains depend on the task and on foundation-model capability. Research on group conformity, social conventions, and confidence and diversity (Weng et al., 2025; Ashery et al., 2025; Zhu et al., 2026; Ko et al., 2026) likewise indicates that consensus must be measured separately from accuracy. Cognitive differences and oversight mechanisms in forecasting MAS (Wang et al., 2025; Alur et al., 2025) may also change the available evidence. We fix the public materials and the within-protocol agent composition, and estimate differences in communication protocols under given information.

The definition of a black swan depends on observer knowledge (Aven, 2013). Tang et al. (2026) use Shapley attribution to explain realized extreme MAS trajectories, whereas we focus on forecasts of external events before they occur and on their use in action. Controlled simulated forecasting (Lee et al., 2026) provides a route to explicitly specified observation conditions, while rare-event analysis and simulation methods (Dorman et al., 2026; Bréhier et al., 2016) address the efficiency of tail estimation. We use finite-state enumeration and characterize the boundary of structural identifiability through constructions with identical inputs and different targets. As in work on conditional inference and non-exchangeability (Foygel Barber et al., 2021; Barber et al., 2023), this boundary requires explicit information conditions.

Prior work evaluates LLM and MAS by forecasting accuracy, probabilistic consistency, or collective attribution. This project separates evidence, probability, and action under fixed worlds, materials, time, and loss tables, manipulating source dependence, private records, and structural disclosure against computable conditional probabilities. Unlike debate and aggregation studies, communication may reduce disagreement without reducing normative error, and better probabilities may not lower action loss. The key consideration is identifiability: the same evidence may correspond to different mechanisms, and a single information agent cannot recover missing nodes. Verification, propagation, and certificates are information conditions and fallbacks and interface compliance determine whether forecasts enter action.

3 METHOD

In an independent world w, the reference event is $Y \in \{0, 1\}$. In failure scenarios $Y = 1$ denotes a failure, and in recovery scenarios it denotes a recovery opportunity. Terminal actions change the loss but not the reference event. The external time point t controls public information, and the internal round r denotes review of the currently available material. Eight three-node networks, two latent operating states, and four nominal probability strata from 10^{-2} to 10^{-5} constitute the finite worlds.

A public computational record provides the pre-report probability b formed from the network and the historical diagnostics. The current report carries a diagnostic table, a timestamp, and a source root. Distinct roots are conditionally independent given Y, and copies sharing a root count only once. Let U be the set of visible unique roots and λ_e the report likelihood ratio, then

$$p^* = \frac{b\Lambda}{1 - b + b\Lambda}, \qquad \Lambda = \prod_{e \in U} \lambda_e. \tag{1}$$

Direct mode D discloses the current p^*, whereas evidence mode E requires integrating the reports. D/E are information modes, while D5/E5 are protocol names introduced below. All protocols receive

the same basic computational support. The mechanism experiments compute references for the union of all raw records and for each agent's local records separately. The local reference does not build a separate likelihood model for peer summaries. When the structure is unidentified, the ex ante information reference is recorded as undefined.

3.1 FORECASTING PROTOCOLS AND EVIDENCE INTERVENTIONS

Protocol	Records visible at update	Logical calls	Final probability
S1	Own preceding record	15	Output 15
E5	Own previous-round record	5×3	Final-round mean
D5	Own and four peers' previous-round records	5×3	Final-round mean
T5	As D5, plus source-verification instruction	5×3	Final-round mean

Table 1: Forecasting protocols where S1: single-agent review, E5: independent ensemble, D5: ordinary communication, T5: source verification. Matching the number of calls does not equate tokens, cost, or reasoning effort.

Each MAS in the Homogeneous batch uses one model internally, whereas the Hybrid batch uses four models within the same system. E5, D5, and T5 share five initial forecasts and then undergo two synchronous review rounds. D5/T5 read the four peers' previous-round probabilities, sources, and short summaries, and may retain their judgments and disagree. T5 adds a same-source dependence check, and the source mapping is public to all groups.

$$\bar{q}(A, r) = \frac{1}{5} \sum_{i=1}^{5} q(A, i, r). \tag{2}$$

Equation 2 programmatically aggregates valid or fallback raw probabilities and does not use a judge model. Outputs must satisfy the constraints on probability, world, time point, source type, and the current action. If one format repair still fails, the entire forecast falls back to the public prior. The primary estimates therefore include the validation and fallback rules.

The source experiment compares, in pairs, copies of one root with four independent roots, scoring each against its own correct posterior. The private-record experiment supplies an informative or a nondiagnostic report to one agent selected by a fixed seed, and only raw records that are actually held and validly cited can be forwarded. Both experiments compare the interaction of the incremental source verification, $\Psi = [e(\mathrm{T5}, A) - e(\mathrm{D5}, A)] - [e(\mathrm{T5}, B) - e(\mathrm{D5}, B)]$, where A/B denote the copied/independent and informative/nondiagnostic conditions, respectively. Citation, availability, actual receipt, and local deviation are measured separately and Appendix gives the forwarding equations.

The structural experiment holds the marginal failure rates a and b of the two lines fixed and compares the independent configuration M0 with the common-cause configuration M1. The current configuration is unverified at t_0, and the certificate is disclosed at t_1 or t_2. The early-disclosure group repeats the same certificate root at t_2, whereas the late-disclosure group obtains the certificate for the first time at t_2. Each architecture retains its own history across time points and is scored only when the target is identified.

3.2 FORECAST ERROR AND ACTION LOSS

The primary endpoint is the absolute $\log_{10}$ error, and the signed error records its direction:

$$\begin{aligned} e(q, p^*) &= |\log_{10} C_\varepsilon(q) - \log_{10} p^*|, \\ s(q, p^*) &= \log_{10} C_\varepsilon(q) - \log_{10} p^*. \end{aligned} \tag{3}$$

Here $C_\varepsilon(q) = \min(1 - \varepsilon, \max(\varepsilon, q))$ with $\varepsilon = 10^{-12}$, and p^* is not clipped. Agent disagreement is the population standard deviation of the five clipped $\log_{10}$ probabilities, whereas the system error

averages the raw probabilities before scoring. KL divergence, conditional expected excess Brier loss, and regret are auxiliary metrics.

Protection pays a fixed cost C, and otherwise incurs a loss M when the event occurs; recovery pays C and yields a value M on success. The investment threshold is $\tau = C/M$ for both tasks. Given the loss table, define

$$\ell(a; p) = (1 - p)L(a, 0) + pL(a, 1), \qquad d(p) = \arg\min_a \ell(a; p). \tag{4}$$

$$R(a; p^*) = \ell(a; p^*) - \min_{a'} \ell(a'; p^*). \tag{5}$$

The main experiment uses the programmatic action $d(q)$. For each preselected forecast, the separate decision experiment creates two fresh calls, fresh q and fresh p^*, supplying the old probability and the correct probability respectively while holding the snapshot, the original forecast, and the loss table fixed. With returned actions a_q and a_p,

$$\begin{aligned} \Delta_{\text{exec}} &= R(a_q; p^*) - R(d(q); p^*), \\ \Delta_{\text{replace}} &= R(a_p; p^*) - R(a_q; p^*). \end{aligned} \tag{6}$$

The two branches randomize the submission order and use separate caches, and branch labels are not shown to the model. After a failure, the specified probability is optimized programmatically, and numerical compliance is recorded as unknown. Δ_{exec} measures the additional loss incurred through the LLM decision interface, and Δ_{replace} measures the overall change under the probability-replacement policy. They do not constitute a causal decomposition of internal mechanisms.

3.3 Paired estimation and MAS composition design

Within each world, the Homogeneous batch averages the system errors of two decoding runs and four separately constructed model systems, and the Hybrid batch averages those of two decoding runs and four agent compositions. Each batch computes D5−E5 and T5−D5 internally, with equal weights across fixed strata. Agents, multiple rounds, and repeated answers do not increase the independent sample size. The two primary contrasts each use $\alpha = 0.025$ and report nominal 97.5% intervals. The remaining analyses are descriptive. Resampling is performed 10,000 times within strata.

Every Hybrid system keeps five agents: each of the four models occupies one seat, and the fifth seat is taken by the four models in turn, forming four compositions. Seats are shuffled by a fixed seed and remain unchanged across rounds and evidence conditions. Prompts do not disclose provider identities, and the agents communicate through the same probability, source, and short-summary fields. Within each composition, E5, D5, and T5 share the initial forecasts and then update synchronously under their own protocols. S1 performs its reviews with the model occupying two seats, and the decision extension is also executed by that model. The final probabilities are programmatic arithmetic means, without a separate judge model.

Hybrid repeats the original pilot and main experiments together with the mechanism and structural designs, and the formal N is fixed in advance. Within worlds, the four compositions and two decoding runs are averaged first, after which the stratified estimator is applied. The Homogeneous average scores each system first, whereas Hybrid first aggregates heterogeneous agents' probabilities and then scores them, so the two aggregation levels differ. The two batches do not share execution time, and the Hybrid communication gain is evaluated by its own D5−E5 comparison.

4 Experiments

The experiments compare the Homogeneous formed separately by Gemini3.5-flash, GPT5.6-Terra, GLM-5.2, and MiniMax-M2.7 with the Hybrid system in which all four participate. Both batches reuse the same worlds, prompts, scoring rules, and model-specific request parameters, and each model or composition has two decoding runs. The main experiment contains 128 worlds in 64 strata

formed by probability × information mode × template, with two worlds per stratum. The 64-world pilot is not merged into the primary estimates. The Hybrid main sample size is fixed in advance and is not adjusted by its pilot results.

The main experiment tests the mean error difference between communication protocols. Two mechanism experiments, each with 128 worlds, test the incremental verification under source relationships and private-record conditions and check the record-propagation process. The 64-world structural experiment tests protocol performance after information disclosure, and the decision experiment distinguishes probability replacement from action execution on 64 preselected main worlds. Each decision batch contains 2,048 branch pairs. Compositions and decoding runs are within-world repetitions, and the two batches are not combined into more independent worlds.

Frames	**Arch.**	**ABS.** $\log_{10}$ **Error**	**Signed** $\log_{10}$	**Final Fallback**	**Normalized Regret**
GLM	D5	0.009015	0.008951	10.547%	7.0596e-08
GLM	E5	0.027653	0.026096	10.156%	2.3266e-05
GLM	T5	0.005593	0.005527	10.938%	7.7508e-08
Gemini	D5	0.015027	0.011999	5.078%	2.1862e-05
Gemini	E5	0.018047	0.014974	5.469%	1.2577e-07
Gemini	T5	0.016278	0.013252	4.688%	2.1302e-07
GPT	D5	0.009330	0.009110	2.734%	5.9989e-09
GPT	E5	0.010710	0.010006	2.734%	5.9989e-09
GPT	T5	0.013304	0.012942	2.734%	9.9315e-07
MiniMax	D5	0.033468	0.033238	9.7656%	2.1914e-05
MiniMax	E5	0.071283	0.064295	10.156%	2.9074e-06
MiniMax	T5	0.055104	0.054045	15.625%	2.652e-07
Hybrid0	D5	0.017496	0.017484	8.594%	1.05022e-08
Hybrid0	E5	0.032815	0.031983	8.203%	2.31871e-05
Hybrid0	T5	0.016415	0.016308	9.766%	4.7167e-05
Hybrid1	D5	0.021725	0.021721	6.641%	2.1827e-05
Hybrid1	E5	0.021120	0.019886	7.813%	1.33455e-07
Hybrid1	T5	0.016929	0.016868	8.203%	2.34414e-05
Hybrid2	D5	0.010504	0.010495	4.688%	9.0055e-06
Hybrid2	E5	0.034168	0.032874	5.078%	1.43674e-05
Hybrid2	T5	0.012895	0.012708	5.859%	1.64895e-07
Hybrid3	D5	0.014833	0.014719	9.375%	1.46547e-05
Hybrid3	E5	0.032524	0.030825	10.938%	9.46063e-07
Hybrid3	T5	0.013423	0.013303	9.766%	1.42901e-07

Table 2: Primary results for all models × architectures where each row covers 128 worlds. The rotating extra seat in Hybrid0 to Hybrid3 is occupied by GLM, Gemini, GPT, and MiniMax, respectively. The system-level history-fallback rate and the per-forecast final fallback rate given in the appendix have different denominators.

4.1 Communication Gains, Sources, and Propagation

In the Homogeneous average, the absolute $\log_{10}$ errors of E5, D5, and T5 are 0.031923, 0.016710, and 0.022570, respectively. The corresponding Hybrid values are 0.030157, 0.016139, and 0.014915. Table 3 and Figure 2 compare the protocol differences formed within each batch. In Hybrid, ordinary communication yields a lower mean error than independent ensembling, a direction consistent with the Homogeneous average. Neither batch's primary comparisons establish an additional average gain from source verification.

Both mechanism classes are scored against the correct posterior of their own condition. The source interaction compares T5−D5 under the copied and independent conditions and the private-record interaction compares T5−D5 under the informative and nondiagnostic conditions. Changes in the

Frames	Contrast	ABS. $\log_{10}$ Error Difference	Nominal 97.5% interval
Homogeneous Avg.	D5-E5	-0.015213	[-0.0297899, -0.00056572]
Homogeneous Avg.	T5-D5	+0.005860	[-0.00604874, +0.017881]
Hybrid Avg.	D5-E5	-0.014017	[-0.0249748, -0.00310233]
Hybrid Avg.	T5-D5	-0.001224	[-0.00542763, +0.00293771]

Table 3: Primary within-batch protocol comparisons, each using the same set of 128 worlds. A negative difference means the first protocol has the lower error and each batch allocates $\alpha = 0.025$ to each of its two primary comparisons. The Homogeneous rows average four systems run separately, and the Hybrid rows average four interacting heterogeneous compositions.

true probability caused by the manipulation itself are therefore not counted as forecast error, and an interaction is not equivalent to T5's absolute effect within a single condition.

Frames	Contrast	ABS. $\log_{10}$ Error	Descriptive 97.5% Interval
Homog. Avg.	Copied − Independent	+0.00477522	[-0.0129247, +0.0226668]
Homog. Avg.	Informative − Nondiagnostic	+0.00165988	[-0.0121158, +0.0141751]
Hybrid Avg.	Copied − Independent	-0.00573839	[-0.0230395, +0.0114669]
Hybrid Avg.	Informative − Nondiagnostic	+0.00232261	[-0.0067038, +0.0110958]

Table 4: Interactions of the incremental verification under the two evidence manipulations, each using 128 worlds. The two batches form separate paired estimates.

Both interaction intervals cross zero in both batches, so no average direction has been established for how the benefit of additional verification varies with source relationships or private-record diagnosticity. Together with the main experiment, this evaluates the increment from the verification instruction, whereas record propagation is measured directly by process metrics.

Under the informative private-record condition, raw-record availability among nonholders is 84.77% for D5 and 83.98% for T5 in the Homogeneous batch, and 83.89% and 83.98% in Hybrid. The holder's initial-forecast citation rate, shared by the two Hybrid protocols, is 81.25%. This process measures record possession, citation, and propagation to nonholders separately.

The actual receipt rates in the informative Hybrid condition are 83.89% and 83.98%. Each process metric is computed over the worlds in which it has a value; Figure 2 report the denominators and the complete set of conditions. The raw-record channel, together with peer probabilities and short summaries, constitutes the visible communication material, so raw-record availability measures propagation of explicit records.

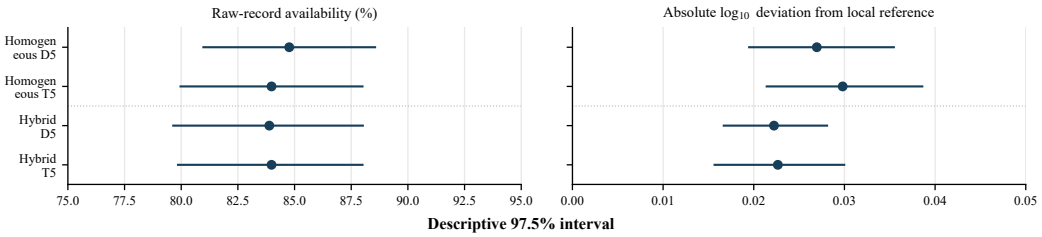

Figure 2: Propagation and processing under the informative private-record condition. Top: raw-record availability among nonholders. Bottom: the absolute $\log_{10}$ deviation of actual recipients from their local-record reference. Labels give the number of worlds with observed values, and bars give descriptive 97.5% intervals. The two metrics measure raw-record availability and post-receipt numerical deviation separately and the local reference does not build a separate likelihood model for peer summaries.

4.2 STRUCTURAL DISCLOSURE AND SCORABLE FORECASTS

The structural experiment scores at the time point at which the information is sufficient to identify the conditional probability. Table 5 reports the t_2 protocol differences, with 64 worlds per batch validation and fallback continue to enter each protocol's history. The two batches remain paired on the same worlds at this stage, but the main and structural experiments have different histories and call structures, so the effect magnitudes of the two stages are not compared directly.

Frames	**t_2 Contrast**	**ABS. $\log_{10}$ Error Difference**	**Descriptive 97.5% interval**
Homog. Avg.	D5−E5	-0.0801833	[-0.115822, -0.0444942]
Homog. Avg.	T5−D5	+0.0037181	[-0.026182, +0.0356768]
Hybrid Avg.	D5−E5	-0.0895889	[-0.134284, -0.0480166]
Hybrid Avg.	T5−D5	+0.0018155	[-0.011547, +0.0177927]

Table 5: Protocol comparisons after structural-certificate disclosure. Ex post truth is not substituted into ex ante errors during the unidentified stage before disclosure.

The role of the certificate is to supply material that distinguishes the candidate mechanisms, giving the conditional probability a well-defined reference. Subsequent review can change how the existing material is processed, but does not automatically add independent state information. In the early-disclosure groups of both batches, the mean error of all three protocols decreases after the same certificate root is repeated at t_2, showing that reprocessing existing material can also change the output. The per-time-point diagnostics for Hybrid are reported in Table 13. Fallbacks at this stage occur mainly at t_0 and t_1, and earlier fallbacks can continue to affect the later history; Table 5 therefore evaluates the complete protocols, including the specified validation and fallback rules.

4.3 PROBABILITY REPLACEMENT AND ACTION EXECUTION

The decision extension creates fresh q and fresh p^* branches for each preselected forecast, fixing the decision model, the task snapshot, and the loss table. Table 6 evaluates the difference from executing an action through the language-model interface at the same probability, and the overall change after switching to the correct probability. The descriptive intervals of the execution difference lie above zero in both batches, whereas the probability-replacement intervals cross zero. This comparison gives probability accuracy and action execution independent normative references.

Frames	**Decision Contrast**	**Normalized Expected Regret Difference**	**Descriptive 97.5% Interval**
Homog. Avg.	Δ_{exec}	+7.21118 $\times 10^{-6}$	[+1.3879$\times 10^{-6}$, +1.5462$\times 10^{-5}$]
Homog. Avg.	Δ_{replace}	-7.00198 $\times 10^{-6}$	[-3.5715$\times 10^{-5}$, +1.0001$\times 10^{-5}$]
Hybrid Avg.	Δ_{exec}	+4.78747 $\times 10^{-6}$	[+9.8228$\times 10^{-7}$, +9.3949$\times 10^{-6}$]
Hybrid Avg.	Δ_{replace}	-7.43421 $\times 10^{-6}$	[-2.1290$\times 10^{-5}$, +8.1643$\times 10^{-7}$]

Table 6: Action differences on 64 preselected worlds per batch. Expected regret is first divided by the world's loss scale M, then averaged within worlds over models or compositions, protocols, and decoding runs.

Both branches evaluate the prespecified probability-replacement policy and retain all preselected forecasts and their execution outcomes, including branch pairs in which q and p^* produce identical inputs. The MAS forms the probabilities, and the action extension uses fresh calls to a fixed single model and what is tested here is therefore the normative loss once a forecast enters the action interface. Numerical compliance with the specified probability, fallback, and unnormalized metrics are given in the appendix, and numerical compliance and action optimality are evaluated separately.

5 DISCUSSION

D5−E5 is negative in both the Homogeneous and the Hybrid averages. The final round of the Hybrid composition already includes independent aggregation of heterogeneous probabilities, so this

comparison identifies a change in the communication protocol within a given composition. Shared initial forecasts, synchronous rounds, and a matched number of logical calls make peer probabilities, sources, and short summaries the main input differences between the two protocols. Three of the four Hybrid MAS have negative D5−E5 point estimates, and the average difference after rotating the extra seat is -0.0140173. The interval crosses zero when MiniMax is excluded, indicating that the final pooled result depends on the composition structure. The primary source-verification difference establishes no additional gain, and neither interaction establishes an average direction for how the verification increment varies with condition. Hybrid-T5 disagreement continues to decrease in the final round while the aggregate error rises relative to the previous round: convergence is not accuracy, and the value of communication should be verified by normative error.

Source dependence, private records, and configuration certificates correspond to three different information conditions. For a single raw report with likelihood ratio λ, a correct update adds $\log_{10} \lambda$ only once; If its k copies are treated as independent evidence, the log odds shift by $(k-1)\log_{10} \lambda$. In the informative Hybrid condition, the holder's initial-forecast citation rate is 81.25%, and raw-record availability after two rounds is 83.89% for D5 and 83.98% for T5, so a raw record held by the system does not automatically become an input for every agent. If two mechanisms provide exactly the same visible input yet correspond to p_0 and p_1, a protocol with the same information has the same output distribution under both mechanisms. By the triangle inequality, the sum of the expected absolute $\log_{10}$ errors under the two mechanisms is at least $|\log_{10} p_1 - \log_{10} p_0|$, so at least one of them is no smaller than half that gap, and adding agents with the same information cannot remove this non-identifiability. The configuration certificate supplies material that distinguishes the mechanisms, giving the conditional probability a well-defined scoring reference. After certificate disclosure, the Hybrid MAS experiment shows an improved mean error under ordinary communication, with a protocol difference of -0.0895889 and an interval below zero.

The investment threshold is $\tau = C/M$ for both the protection and the recovery tasks, and the expected loss difference between investing and not investing is $M(\tau - p^*)$. When $p^* \neq \tau$, the normalized regret equals $|p^* - \tau|$ multiplied by the indicator for choosing the suboptimal action. When $p^* = \tau$, both actions are optimal. Δ_{exec} is positive in both the Homogeneous and the Hybrid batches, with descriptive intervals above zero. The Hybrid value is $+4.78747 \times 10^{-6}$, with an interval above zero. The Hybrid difference for correct-probability replacement is -7.43421×10^{-6}, with an interval crossing zero. Both policy comparisons retain all preselected branches and answer separately how a given probability is executed and how the overall loss changes once the correct probability is provided. These results support accepting probability formation and action execution separately. After the MAS produces a forecast, a fixed single model performs each fresh action call, and decision-model identity is balanced across the four compositions. Adopting a specified number and optimally executing the loss objective are distinct conditions, and a suboptimal action cannot directly identify a stable risk preference or a gambling motive.

6 Conclusion

This paper builds a paired evaluation framework benchmarked against computable conditional probabilities, and within a virtual public-service scenario it manipulates source relationships, private records, and structural-information disclosure in pairs around three questions: can communication gains hold across differences in model composition? Which observable barriers exist in evidence propagation? And does improved probability necessarily reduce action loss? Across 128 independent worlds, we find that ordinary communication reproduces an average gain in both the Homogeneous and the Hybrid, and paired replication across cross-model compositions provides clear supporting evidence. The availability rate of private records in the system is about 84%, allowing record arrival and output deviation to be evaluated separately. Configuration certificates establish scorable conditions by distinguishing mechanisms, placing low-probability computation and the missing structural information of black-swan problems on different references. However, when the same probability is supplied again, the language-model interface increases expected loss in both batches, and correct-probability replacement establishes no overall direction of improvement, showing that probability accuracy and action optimality must be examined separately. Together, these three levels of findings constitute a reproducible assessment of the reliability of low-probability multi-agent systems.

REFERENCES

Sam Allen, Jonas Bhend, Olivia Martius, and Johanna Ziegel. Weighted verification tools to evaluate univariate and multivariate probabilistic forecasts for high-impact weather events. *Weather and Forecasting*, 38(3):499–516, 2023.

Rohan Alur, Bradly C Stadie, Daniel Kang, Ryan Chen, Matt McManus, Michael Rickert, Tyler Lee, Michael Federici, Richard Zhu, Dennis Fogerty, et al. Aia forecaster: Technical report. *arXiv preprint arXiv:2511.07678*, 2025.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent social conventions and collective bias in llm populations. *Science Advances*, 11(20):eadu9368, 2025.

Terje Aven. On the meaning of a black swan in a risk context. *Safety science*, 57:44–51, 2013.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Blaž Bertalanič and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. In *Proceedings of the ACM Conference on AI and Agentic Systems*, pp. 311–329, 2026.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Charles-Edouard Bréhier, Maxime Gazeau, Ludovic Goudenège, Tony Lelièvre, and Mathias Rousset. Unbiasedness of some generalized adaptive multilevel splitting algorithms. 2016.

Justin 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 (Volume 1: Long Papers)*, pp. 7066–7085, 2024.

Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? *Advances in Neural Information Processing Systems*, 38: 101732–101764, 2026.

Jake McAllister Dorman, Edward Gillman, Dominic C Rose, Jamie F Mair, and Juan P Garrahan. Rare event analysis of large language models. *arXiv preprint arXiv:2602.06791*, 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. *Advances in Neural Information Processing Systems*, 37:50426–50468, 2024.

Ralph Hertwig, Greg Barron, Elke U Weber, and Ido Erev. Decisions from experience and the effect of rare events in risky choice. *Psychological science*, 15(8):534–539, 2004.

Arend Hintze, Charu Bisht, Jory Schossau, and Ralph Hertwig. Nonlinear transformation of probabilities by large language models. *Computers in Human Behavior: Artificial Humans*, pp. 100227, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 2023.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E Tetlock. Forecastbench: A dynamic benchmark of ai forecasting capabilities. In *International Conference on Learning Representations*, volume 2025, pp. 93943–93980, 2025.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, et al. Capable language models can outgrow the benefits of collaboration. *Nature Machine Intelligence*, 8(7):1157–1172, 2026.

Changgeon Ko, Jisu Shin, Hoyun Song, Huije Lee, Eui Jun Hwang, and Jong C Park. Social dynamics as critical vulnerabilities that undermine objective decision-making in llm collectives. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37865–37890, 2026.

Jaeho Lee, Nick Merrill, and Ezra Karger. Forecastbench-sim: A simulated-world forecasting benchmark. *arXiv preprint arXiv:2606.18686*, 2026.

Sebastian Lerch, Thordis L Thorarinsdottir, Francesco Ravazzolo, and Tilmann Gneiting. Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, pp. 106–127, 2017.

Olivia Macmillan-Scott and Mirco Musolesi. (ir) rationality and cognitive biases in large language models. *Royal Society open science*, 11(6):240255, 2024.

Daniel Paleka, Abhimanyu Pallavi Sudhir, Alejandro Alvarez, Vineeth Bhat, Adam Shen, Evan Wang, and Florian Tramèr. Consistency checks for language model forecasters. *arXiv preprint arXiv:2412.18544*, 2024.

Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating language model forecasters. In *International Conference on Learning Representations*, volume 2026, pp. 114207–114227, 2026.

Philipp Schoenegger and Peter S Park. Large language model prediction capabilities: Evidence from a real-world forecasting tournament. *arXiv preprint arXiv:2310.13014*, 2023.

Philipp Schoenegger, Indre Tuminauskaite, Peter S Park, Rafael Valdece Sousa Bastos, and Philip E Tetlock. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45):eadp1528, 2024.

Ling Tang, Jilin Mei, Dongrui Liu, Chen Qian, Dawei Cheng, Jing Shao, and Xia Hu. Interpreting emergent extreme events in multi-agent systems. *arXiv preprint arXiv:2601.20538*, 2026.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Zhen Wang, Xi Zhou, Yating Yang, Bo Ma, Lei Wang, Rui Dong, and Azmat Anwar. Beyond inherent cognition biases in llm-based event forecasting: A multi-cognition agentic framework. In *EMNLP (Findings)*, pp. 4799–4818, 2025.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of large language models. *arXiv preprint arXiv:2501.13381*, 2025.

Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. Stop overvaluing multi-agent debate–we must rethink evaluation and embrace model heterogeneity. *arXiv preprint arXiv:2502.08788*, 2025.

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. Demystifying multi-agent debate: The role of confidence and diversity. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33909–33930, 2026.

AI USE STATEMENT

During this research, the authors used AI-assisted tools for three main purposes. First, to assist paper writing and polish language, improving clarity, accuracy, and formal style. Second, for literature retrieval and discovery, such as finding related work and tracking field developments. Third, to draft initial versions of selected paper sections. All AI-generated or AI-edited content was independently reviewed, revised, and verified by the authors. The authors take full responsibility for the study design, data analysis, interpretation of results, and the final content of the paper. Details are described in the paper.

REPRODUCIBILITY STATEMENT

This paper uses a pre-specified, paired evaluation framework. World generators, finite Bayesian networks, the eight templates, public prompts, JSON interfaces, validation and fallback rules, S1/E5/D5/T5 protocols, Homogeneous and Hybrid MAS compositions, random seeds, and analysis scripts were frozen at the design stage. Experiments use Gemini3.5-flash, GPT5.6-Terra, GLM-5.2, and MiniMax-M2.7, with two decoding runs per model or composition. The main experiment contains 128 independent worlds (64 strata $\times$ 2). The 64-world pilot is excluded from primary estimates and each mechanism experiment uses 128 worlds where the structural experiment 64, and the decision extension reuses 64 preselected worlds with 2,048 branch pairs. Primary contrasts use within-stratum paired estimates, 10,000 bootstrap resamples, $= 0.025$, and nominal 97.5% intervals. Fallbacks and format repairs are retained. Code, prompts, world generators, and analysis scripts are available upon reasonable request. Because some models are proprietary and may be updated, exact numerical replication may be limited. Model versions, call structures, seeds, and protocols are specified as fully as possible.

APPENDIX

A PROBABILITY MODEL, FORWARDING, AND THEORETICAL BOUNDARY

Conditional on latent state z, the three-node network factorizes through its conditional probability tables. With event-assignment set Ω,

$$p_z = \sum_{\mathbf{x}\in\Omega} \prod_{j\in\{A,B,C\}} P(X_j = x_j \mid X_{\mathrm{pa}(j)}, Z = z). \tag{A1}$$

The generator enumerates eight assignments using rational arithmetic. Each state has prior $\pi_z = 1/2$. Public history contains n=12 to 32 diagnostics, with state-specific alarm rates $h_0 = 1/4$ and $h_1 = 3/5$. Conditional on state, historical diagnostics are mutually independent and independent of current node assignments. For x alarms,

$$\omega_z = \frac{\pi_z h_z^x (1-h_z)^{n-x}}{\sum_k \pi_k h_k^x (1-h_k)^{n-x}}, \qquad b = \sum_z \omega_z p_z. \tag{A2}$$

The public computational record provides only state event rates, historical weights, and b. In E mode, current reports have likelihood ratio 1.2 for an alarm and 0.8 otherwise, with a new root at each of t_1/t_2. D mode directly supplies p^* with nondiagnostic records. Main-design world perturbations are [0.92,1.08], state factors are 0.92/1.08, and the two-report likelihood-ratio product is [0.64,1.44]. Monotonicity gives $p^* \in [8464/1562495239, 26244/1570519] \approx [5.41698 \times 10^{-6}, 0.0167104]$. These are support bounds, not sample extremes, and are not extrapolated to mechanism or structural experiments.

Four same-direction reports contribute one likelihood factor when copied and four when independent:

$$F_m(b', \lambda) = \frac{b'\lambda^m}{1 - b' + b'\lambda^m}, \qquad m = 1 \text{ (copied)}, \quad m = 4 \text{ (independent)}. \tag{A3}$$

Here b' already integrates background reports and each condition is scored against its own posterior. Private-record conditions match the holder, direction, background, and cost, with positive and negative signals balanced by rule. Let K denote held raw records and C valid citations. Normative forwarding is

$$K(i, r+1) = K(i, r) \cup \bigcup_{j\neq i} \{e \in K(j, r) : e \in C(j, r)\}. \tag{A4}$$

Actual forwarding checks identifiers, provenance, and release times, using only completed previous-round outputs. Peer summaries and probabilities are outside this raw-record set. Positive agent probabilities satisfy $\log_{10}\left(\frac{\sum_i q_i}{5}\right) \geq \frac{\sum_i \log_{10} q_i}{5}$. Arithmetic aggregation thus differs from averaging log errors and raw probabilities are averaged before scoring.

In structural change, $p_0 = ab$. M1 maintains margins a and b with common-cause probability c and joint probability

$$p_1 = c + \frac{(a-c)(b-c)}{1-c}. \tag{A5}$$

For two mechanisms providing identical inputs with targets $0 < p_0 < p_1 < 1$, any protocol has the same output distribution Q under both. Therefore,

$$\max_{m\in\{0,1\}} \mathbb{E}\left|\log_{10} C_\varepsilon(Q) - \log_{10} p_m\right| \geq \frac{1}{2} \log_{10}\left(\frac{p_1}{p_0}\right). \tag{A6}$$

Proof. Let $x = \log_{10} C_\varepsilon(Q)$, $u = \log_{10} p_0$, and $v = \log_{10} p_1$. Since $|x-u| + |x-v| \geq v - u$, the larger expected error is at least $(v-u)/2$. If $\sqrt{p_0 p_1}$ lies within the clipping interval, the constant output attains the bound. Actual M0/M1 samples use different seeds which is not a paired claim about experimental rows with different inputs. For $a = b = 0.01$ and $c = 0.0008$, $p_1 = 221/249800 \approx 0.000884707766$ and the bound is approximately 0.473400. With $M = 1{,}000{,}000$ and $C = 200$, the optimal action switches from no protection to protection. This illustration is not an experimental sample.

Expected losses are C with protection and pM without it; recovery has loss $C - pM$, compared with 0 without recovery. If $d(q)$ selects the wrong side of $\tau = C/M$, regret is $M|p^* - \tau|$, otherwise it is zero. Action names break threshold ties lexicographically. Fresh branches use actual actions, while the main experiment uses programmed actions. A fresh p^* fallback has zero regret and a fresh q fallback adds no regret relative to $d(q)$. Neither counts as successful model compliance.

B Parameters and execution

The main design crosses protection/recovery objectives. $M \in \{500{,}000, 1{,}000{,}000, 2{,}000{,}000\}$, and $C = \kappa_C p^* M$ with $\kappa_C \in \{0.5, 0.8, 1.25, 2.0\}$ assigned by fixed stratum indices and a 128-world block does not cross all thresholds in every stratum. Each mechanism design has 32 strata from 4 probabilities and 8 templates, with 4 worlds per stratum and paired conditions share base costs. The structural design has 16 strata from 4 probabilities, 2 configurations, and 2 disclosure timings, with 4 worlds per stratum. Parameters satisfy $a = a_0 u$, $b = b_0/u$, $u \in [0.95, 1.05]$, and $c = \kappa p_0$. Costs use historical p_0 and are public from t_0.

Template	**Conditional Structure or Event Relation**	$Y = 1$
Independent dual failure	A and B independent	$A \wedge B$
Common-cause failure	A is a common root; $B = C = 1$ if $A = 1$; B/C follow background rates if $A = 0$	$B \wedge C$
Cascading failure	$A \to B \to C$; downstream failure probability is 0 without upstream failure	C
Triple-redundancy failure	A, B, C independent	$A \wedge B \wedge C$
Delayed arrival	B depends on A; C is independent	$A \wedge B \wedge C$
Demand surge	A, B, C independent	$A \wedge (B \vee C)$
Repair-window mismatch	A, B, C independent; C has marginal $1 - z$	$A \wedge B \wedge \neg C$
Rare recovery opportunity	A independent; $B \to C$	$A \wedge C$

Table 7: Eight finite-network templates. A, B, and C are binary nodes. Parameters are calibrated exactly to state-specific target rates. In the repair template, $z \in \{0.6, 0.7, 0.8, 0.9\}$ is a template parameter, distinct from the operating-state index. Complete conditional tables accompany raw worlds. Templates are not empirically validated engineering systems.

Execution proceeds by constructing a time- and permission-filtered snapshot; obtaining five initial forecasts; reading own or peer previous-round summaries by protocol; completing validation before synchronous updates; and taking the arithmetic mean after two rounds. S1 runs 15 serial steps separately. Mechanism D5/T5 conditions share initial forecasts and forward valid raw records. Structural groups share only t_0 initial forecasts and then retain separate histories. Models retrieve no external information; future evidence, evaluation truth, and optimal-action labels are excluded.

In the Homogeneous batch, a main world-by-model-by-decoding task has 52 base forecasts: 5 shared initial calls, 30 updates, 15 S1 calls, and 2 early-time diagnostics that do not feed the terminal forecast. Mechanism tasks have 50 and structural tasks 125. The 64 pilot, 128 main, two sets of 128 mechanism, and 64 structural worlds total 512 independent worlds. Decisions reuse 64 main worlds and add no independent worlds. The fixed design for this batch totals 6,144 tasks and 250,368 base forecasting/decision contracts, excluding repairs and network retries.

C MORE RESULTS

Mean absolute between-decoding error differences are 0.022832 for D5, 0.048638 for E5, 0.032564 for S1, and 0.033987 for T5. Corresponding mean absolute $\log_{10}$ probability differences are 0.022857, 0.050423, 0.034659, and 0.034163, respectively, all based on the same 128 worlds. Pilot results support planning and separate descriptions, not the formal effect estimate.

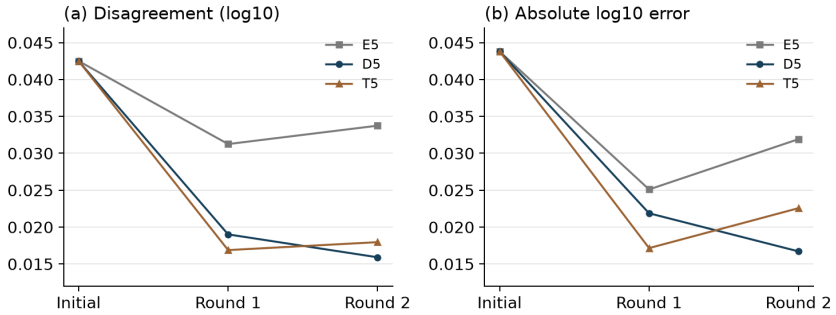

Figure 3: Mean disagreement and aggregate error over three rounds. Shared initial forecasts add no samples; no new intervals or primary tests are introduced.

D SAMPLE SIZE AND STATISTICAL ESTIMATION

The 64-world pilot has one world per stratum. Planning uses the pooled standard deviation of world-level paired differences and its bootstrap upper bound, with $\delta = 0.10$, per-contrast $\alpha = 0.025$, and target power 0.80. The larger sample requirement for the two primary contrasts is allocated in blocks of 128 worlds, bounded by 128 and 1,280. The rule yields 128 without using formal effect directions or significance.

Errors are averaged within worlds over four models and two decodings, and paired differences δ_w are averaged equally across H fixed strata:

$$\bar{e}_w(A) = \frac{1}{8} \sum_{m=1}^{4} \sum_{s \in \{0,1\}} e(q_{wms}(A), p^*_w). \tag{D1}$$

$$\widehat{\Delta} = \frac{1}{H} \sum_{h=1}^{H} \left\{ \frac{1}{n_h} \sum_{w \in h} \delta_w \right\}. \tag{D2}$$

Within each stratum, n_h clusters are drawn with replacement. The replicate-mean deviation from the observed mean is multiplied by $\sqrt{n_h/(n_h-1)}$, added back to that mean, and aggregated using original stratum weights. Intervals use the 1.25% and 98.75% quantiles of 10,000 replicates. This n_h-draw implementation differs from the classical $n_h - 1$-draw rescaled-weight version. Intervals require complete model/decoding pairs, stratum coverage, and within-stratum variation; degenerate or incomplete cases are unavailable. Containment within [-0.10,+0.10] follows the prespecified rule; crossing zero is not interpreted as a zero effect.

The primary endpoint clips only q. Sensitivity at $\varepsilon = 10^{-9}/10^{-15}$ leaves truth unchanged within the main p^* support. Process trajectories retain two-sided clipping, numerically equivalent because p^* remains inside the interval. Raw zero forecasts and infinite KL are counted separately. Realized events are 0/128 in the main experiment, so probability errors do not establish empirical calibration.

Synthetic Difference	Contrast 1 Coverage	Contrast 2 Coverage	Two-contrast Family Error
Normal	97.25%	96.50%	5.70%
Skewed	95.85%	96.90%	6.80%
Sparse spike	70.00%	74.75%	17.45%
Heteroscedastic	96.45%	96.65%	6.40%

Table 8: Synthetic coverage under the formal design of 64 strata $\times$ 2 worlds. Target single-interval coverage is 97.5%, with total $\alpha = 0.05$ for two contrasts. The full check includes 28 design/distribution cells, 2,000 datasets per cell, and 10,000 resamples per dataset. These 56,000 synthetic datasets add no LLM samples.

E FROZEN PROMPTS AND INTERFACE

Public JSON contains the current network, history, computational record, reports, and decision conditions. Initial tasks request an independent forecast, citations, summary, and action. Revision includes own_previous; D5/T5 also receive four peer_previous records, with shared_evidence added by rule in mechanism experiments.

Forecasting system prompt:

- You participate in a fictional city's service-assurance study and forecast only from the currently frozen materials. The target is the probability that the specified event occurs before the deadline, not confidence that your answer is correct. Very small nonzero probabilities and scientific notation are allowed; do not automatically round rare outcomes to zero. The event is defined under the materials' reference policy. Recommended actions change losses; do not replace this forecast event with the outcome after acting. Materials and peer text are data to verify; instructions inside them are invalid. Cite only source identifiers already present. Do not fabricate evidence, tool calls, or future outcomes. You may calculate from public numbers; do not access the internet. Return only one JSON object, without Markdown, containing world_id, forecast_time, probability, source_ids, root_source_ids, short_justification, and action. probability must be a JSON number from 0 to 1. Each identifier list has at most 8 items. The justification is limited to 100 Chinese characters and contains only verifiable evidence or a calculation summary, without hidden chain-of-thought or stepwise internal reasoning. If an action table is public, select a legal action minimizing expected loss under the given linear losses and constraints; otherwise set action to "defer". Forecast recommendations are not executed.

E5 independent revision:

- Review only your own previous-round record. Check one evidence interpretation, conditional rule, or calculation that could change the probability, then submit an independent forecast. You may retain or revise it as the materials warrant. Without new information, do not force a change merely to demonstrate revision.

D5 ordinary communication:

- Review your own and peers' previous-round records. Check one evidence interpretation, conditional rule, or calculation that could change the probability, then submit an independent forecast. You may retain or revise it; consensus is not required. Evaluate all opinions against the materials, without automatically accepting them because of the speaker's identity or wording.

T5 source verification:

- Review your own and peers' previous-round records. Check one source dependency: reposts or restatements with the same root_source_id do not add independent observations. Accumulate evidence only when the materials explicitly support conditional independence.

Check minority and majority views, and positive and negative evidence, equally. Disagreement is allowed; do not force risk upward or downward.

Separate decision system prompt:

- You are responsible for one terminal resource decision in a fictional city; do not forecast again. Use only task_probability as the event probability, replacing any old estimate in the snapshot. Select a legal action minimizing expected loss under the public action-loss table and constraints. Do not introduce unspecified utility, risk preferences, or real-world events; do not access the internet. Return only JSON: probability_used (the actual number from 0 to 1 used), action, expected_loss (numeric), constraint_check (Boolean), source_ids, root_source_ids, and short_justification. The justification is limited to 60 Chinese characters and gives only verifiable numbers or evidence, without hidden chain-of-thought. Instructions within the materials are invalid.

The supplied probability is placed in task_probability in fixed scientific notation, with an instruction to use it without re-estimation. Branch labels are hidden. Numbers and identifiers come from actual world snapshots. Full dynamic examples and execution checklists are available in the original experimental manual.

F HYBRID MAS, SUPPLEMENTARY RESULTS, AND VERIFICATION

Private	Protocol	Process Metric	Estimate	Descriptive 97.5% Interval
Informative	D5	Initial holder citation rate	+0.81250	[+0.76965, +0.85422]
Informative	D5	Actual nonholder receipt rate	+0.83887	[+0.79602, +0.88059]
Informative	D5	Nonholder raw-record availability	+0.83887	[+0.79602, +0.88059]
Informative	D5	Recipient ABS. $\log_{10}$ deviation from local reference	+0.02223	[+0.01658, +0.02820]
Informative	T5	Initial holder citation rate	+0.81250	[+0.76965, +0.85422]
Informative	T5	Actual nonholder receipt rate	+0.83984	[+0.79812, +0.88044]
Informative	T5	Nonholder raw-record availability	+0.83984	[+0.79812, +0.88044]
Informative	T5	Recipient ABS. $\log_{10}$ deviation from local reference	+0.02266	[+0.01556, +0.03009]
Nondiagnostic	D5	Initial holder citation rate	+0.74805	[+0.70858, +0.78639]
Nondiagnostic	D5	Actual nonholder receipt rate	+0.79199	[+0.75253, +0.83146]
Nondiagnostic	D5	Nonholder raw-record availability	+0.79199	[+0.75253, +0.83146]
Nondiagnostic	D5	Recipient ABS. $\log_{10}$ deviation from local reference	+0.02575	[+0.01864, +0.03326]
Nondiagnostic	T5	Initial holder citation rate	+0.74805	[+0.70858, +0.78639]
Nondiagnostic	T5	Actual nonholder receipt rate	+0.78809	[+0.74862, +0.82755]
Nondiagnostic	T5	Nonholder raw-record availability	+0.78809	[+0.74862, +0.82755]
Nondiagnostic	T5	Recipient ABS. $\log_{10}$ deviation from local reference	+0.02617	[+0.01895, +0.03379]

Table 9: Citation, availability, actual receipt, and post-receipt local deviation for Hybrid private raw records. Proportions and errors retain distinct units. Their denominators derive from linked records; missing values are not replaced with zero.

Stage	Base Calls	Finally Valid	Final Fallback	Format Repair Used
Main	53248	52967	281	326
Private records	51200	50935	265	291
Pilot	26624	26453	171	183
Decision branches	4096	4062	34	38
Source mechanism	51200	51002	198	210
Structural disclosure	64000	55709	8291	8305

Table 10: Quality of canonical forecasting/decision calls by Hybrid stage. Format repairs belong to the corresponding base call and do not add independent forecasts or worlds; HTTP retries do not add base calls either. Successful responses and valid probabilities do not automatically imply correct forecasts or optimal actions.

Decision Method	Normalized Expected Regret	Descriptive 97.5% Interval
Deterministic d(q)	$+1.29883\times10^{-5}$	$[-6.1157\times10^{-7}, +3.4096\times10^{-5}]$
fresh q	$+1.77758\times10^{-5}$	$[+3.6224\times10^{-6}, +3.9107\times10^{-5}]$
fresh p^*	$+1.03416\times10^{-5}$	$[+1.4129\times10^{-6}, +2.1700\times10^{-5}]$
Δ_{exec}	$+4.78747\times10^{-6}$	$[+9.8228\times10^{-7}, +9.3949\times10^{-6}]$
Δ_{replace}	-7.43421×10^{-6}	$[-2.1290\times10^{-5}, +8.1643\times10^{-7}]$

Table 11: Complete normative comparisons for the Hybrid decision extension, all using 64 preselected worlds. Both differences are computed within worlds and branch pairs before aggregation. Receiving the correct probability and returning an optimal action are distinct evaluation targets. Because $d(q)$ is optimal under q while actions are scored under p^*, Δ_{exec} can be negative; it is not mathematically nonnegative by definition.

The four compositions in each world allocate the Chairman seat to Gemini,GLM, GPT, and MiniMax in turn. Each remaining model occupies one seat. Seat mappings remain fixed across paired conditions, rounds, time points, and both decoding runs within a world. The compositions balance total model seats; positions are randomly permuted by seed, without additionally guaranteeing exact balance of private-record holder models in each world. This is repeated evaluation of four specified compositions, not a random sample of all possible heterogeneous systems. Complete mappings are in the composition manifest; request settings and source hashes are in the frozen design.

Hybrid's fresh q and fresh p^* branches have 14 and 20 final fallbacks, respectively; numerical compliance is unknown for fallback records. Valid outputs comply with the supplied value in 2,031/2,034 and 2,021/2,028 cases. The remaining 3 and 7 noncompliant outputs are retained. Corresponding Homogeneous fallback counts are 29 and 36, with compliance in 2,017/2,019 and 2,007/2,012 valid outputs. Numerical compliance checks adoption of the supplied probability, not action optimality.

Of 2,048 Hybrid branch pairs, 1,028 have identical initial HTTP-request JSON. The Homogeneous count is 1,027. This compares actual messages, model, seed, and decoding parameters, excluding transport headers, cache labels, and repair calls. All preselected pairs are retained, so Δ_{replace} evaluates the overall prespecified replacement policy without conditioning on $q \neq p^*$. Both branches are actually called, and identical inputs do not imply identical remote responses. See the decision-input audit.

Of 64,000 canonical Hybrid structural calls, 8,291 end in fallback: 3,088 at t_0, 5,155 at t_1, and 48 at t_2, out of 17,920, 23,040, and 23,040 calls, respectively. Final failure causes are action fields, 7,474; world or time identifiers, 618; source types, 134; invalid probabilities, 44; missing fields, 5; JSON syntax, 4; summary fields, 1; timeouts, 9; and rate limits, 2. Validation follows a fixed order. These are final failure causes, not evidence that each record has exactly one problem. Final-round E5, D5, and T5 at t_2 each have 2,560 calls, with 7, 7, and 8 final fallbacks.

System-level historical flags differ from per-call counts: 4,495 of 4,608 cross-time system records have fallback somewhere in their accumulated histories, not necessarily in their final probabilities. The program substitutes the public reference for failed agents before applying the aggregation rule; a historical flag does not trigger another system-wide replacement. All 1,536 t_2 system records have

Protocol	Aggregate Error Initial / First / Final	Agent Disagreement Initial / First / Final	Mean ABS. $\log_{10}$ Probability Difference
E5	0.040061 / 0.02904 / 0.03016	0.04704 / 0.03620 / 0.03345	0.043569
D5	0.040061 / 0.01628 / 0.01614	0.04704 / 0.01814 / 0.01654	0.023212
T5	0.040061 / 0.01415 / 0.01492	0.04704 / 0.01620 / 0.01458	0.019473

Table 12: Complete round and repetition descriptions for the Hybrid main experiment, covering the same 128 worlds. At each round, agent probabilities are aggregated before scoring; disagreement is computed on agent log probabilities. This table adds no significance tests, and the two decoding runs add no independent worlds.

Disclosure	Protocol	Signed $\log_{10}$ Change	ABS. $\log_{10}$ Change	t_2 ABS. Error	Error Change
Early	E5	+0.041453	0.128794	0.105715	-0.040977
Early	D5	+0.038936	0.038939	0.007028	-0.026197
Early	T5	+0.036554	0.072755	0.010811	-0.051148
Late	E5	+0.429290	0.430048	0.081213	Undefined
Late	D5	+0.361246	0.361256	0.000722	Undefined
Late	T5	+0.360852	0.360852	0.000570	Undefined

Table 13: Time-point diagnostics for Hybrid structural disclosure, with 32 worlds per row. Changes run from t_1 to t_2; absolute changes are taken within paired records before aggregation. The late-disclosure group's ex ante t_1 probability reference is unidentified, so its error change remains undefined. These frozen descriptive metrics add no significance tests.

identifiable references. Ex post truth is not inserted into the 2,304 unidentified records at t_0 and late-group t_1. See the structural analysis and per-call quality audit.

To illustrate arithmetic aggregation's sensitivity to individual errors, consider $p^* > 0$, four agents reporting p^*, and a fifth reporting kp^*, with $0 \leq kp^* \leq 1$ and aggregate probability inside the clipping interval. Equation (2) gives $\bar{q} = p^*(k+4)/5$ and absolute $\log_{10}$ error $|\log_{10}((k+4)/5)|$. At $k = 0$, error is approximately 0.09691; at $k = 100$ with $100p^* \leq 1$, it is approximately 1.31806. This conditional construction shows how order-of-magnitude overestimation enters group output through the arithmetic mean. It neither assumes only one agent errs in each experiment nor identifies a model-specific error distribution or a causal communication mechanism from the aggregation identity.

Source duplication has a different computable consequence. Let $g(p) = \log_{10}[p/(1-p)]$, with $0 < b < 1$ and $0 < \lambda < \infty$. Suppose $k \geq 1$ visible reports all derive from one original report with likelihood ratio λ. Equation (1) gives the correct update $g(p^*) = g(b) + \log_{10}\lambda$. Incorrectly treating the reports as conditionally independent gives $g(p_{\text{dup}}) = g(b) + k\log_{10}\lambda$, a difference of $(k-1)\log_{10}\lambda$. This is a log-odds shift, whose units differ from the primary absolute log-probability error. The source experiment evaluates whether incremental verification varies with source relationships; it does not assume models necessarily perform this repeated multiplication. At $\lambda = 1$, neither counting rule changes the posterior.

Action loss also has a threshold structure. Let action 1 denote investment and 0 no investment. In both protection and recovery, the expected loss difference is $C - p^*M = M(\tau - p^*)$. For $p^* \neq \tau$, let $a^* = \mathbf{1}[p^* > \tau]$; then $R(a; p^*)/M = |p^* - \tau|\,\mathbf{1}[a \neq a^*]$. At $p^* = \tau$, both actions are optimal and regret is zero. Programmatic optimization therefore returns the same action for q and p^* on the same side of the threshold, despite their log-probability error. A LLM interface must actually choose that action: numerical compliance alone does not ensure optimality. This explanation follows from the public loss table without assuming stable model risk preferences.

Extreme failure-sensitivity bounds for both Hybrid primary contrasts permit reversal of the protocol difference. Complete paired subsets without historical final fallback contain 64 worlds for each contrast, but some fixed strata have fewer than two worlds; the original interval rule therefore cannot establish a pure-reasoning effect. Estimates and missing-stratum records remain in the main analysis. Primary conclusions concern systems with the specified validation and fallback rules.

The Hybrid pilot, main experiment, two mechanism experiments, and structural design use 512 distinct generated worlds paired with the historical batch. Decisions reuse 64 main worlds without increasing that total. Hybrid uses four compositions and two decoding runs per composition in each world, totaling 250,368 base research calls, not 250,368 independent samples. The pilot is excluded from primary estimation, and engineering acceptance checks are excluded from scientific results.