SCRIP: WHEN DOES A SHARED BUDGET HELP COLLECTIVE REASONING?

Anonymous authors
Paper under double-blind review

ABSTRACT

How should a collective solver allocate a limited budget between acquiring evidence and weighting predictions? We study this question through Scrip, where frozen language agents purchase documents, tools, or peer rationales from the same account that determines their predictive influence. Verified outcomes update account balances during training, and evaluation uses a frozen trained state. Across controlled multi-hop and forecasting tasks, selective acquisition and outcome-based weighting explain most of the improvement over a judge. Coupling acquisition and influence adds a smaller gain at the original budget and a larger gain under scarcity. The advantage is also present under non-oracle retrieval, where retrieval noise narrows it. We characterize the mechanism with two analyses. Matched-report replay identifies sealed Scrip exactly with a cost-adjusted online pool, while a convex-hull bound locates the limit of probability aggregation when reports lack complementary information. These results identify when a shared budget helps collective reasoning and when the task calls for richer communication.

1 INTRODUCTION

Imagine a team of language agents investigating a question whose answer is scattered across several documents. One agent sees a useful abstract, another knows where to find a cross-reference, and a third can check the proposed answer with a tool. Each has enough information to form an opinion, but none necessarily has enough to settle the question. Reading every document and every colleague's reasoning may exceed the available budget. Answering immediately may discard precisely the piece of evidence that would resolve the disagreement. The collective solver must therefore make two decisions: *what should it inspect, and whose judgment should determine the answer?*

This problem arises naturally as language models become components of larger reasoning systems. Sampling and voting combine multiple attempts (Wang et al., 2023); debate and role-based collaboration allow agents to revise their reasoning (Du et al., 2024; Li et al., 2023; Wu et al., 2023); retrieval and tool use extend what each agent can observe (Lewis et al., 2020; Yao et al., 2023). These mechanisms address different parts of collective inference. Yet an aggregation rule operates on the reports available to it, and those reports depend on earlier choices about information access. A reliable agent with no useful new evidence can deserve less additional computation than a less established specialist who knows where the missing fact is located. Historical reliability and the cost of becoming informed must be considered together.

The allocation problem is visible before introducing a market. In our controlled multi-hop setting, equal pooling of 16 agents' reports reaches 62.7% accuracy, yet 61% of its failures contain a correct individual answer. Retrospective label-informed reweighting raises this ceiling further. Closing the gap requires the collective solver to find and credit the useful evidence already present across agents.

An economic perspective suggests a way to study the gap. Markets allocate scarce resources among participants with different information, while outcomes can change their future purchasing power (Hayek, 1945; Kelly, 1956). Recent agent economies use related ideas to choose actions and adapt populations (Qi et al., 2026), and internal LLM markets already aggregate private signals and carry wealth across questions (Galanis, 2026; Begin et al., 2026). We ask a more specific question: *when should the account that pays for information also determine how much a prediction influences the collective answer?*

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

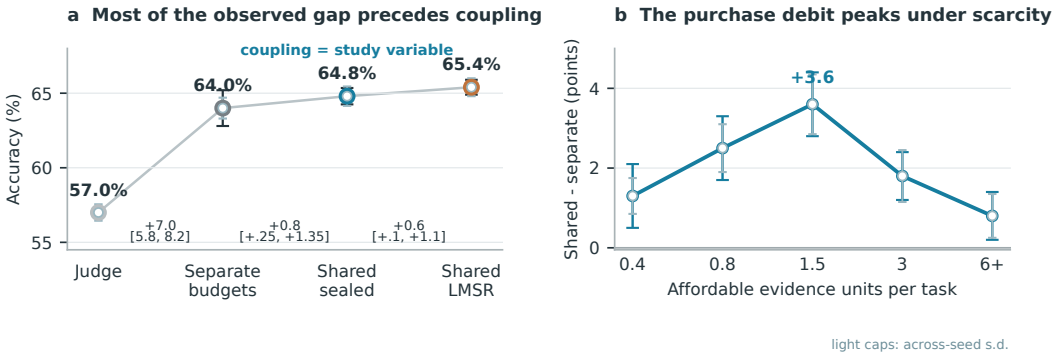

Figure 1: **Evidence access creates room for aggregation, while sharing the account matters most under scarcity. (a)** Controlled 2Wiki configurations for 16 Qwen agents; the shared-versus-separate step is the study variable. **(b)** Sealed shared-minus-separate accuracy across five affordability levels, peaking at $+3.6$ near 1.5 units. Gold-supporting paragraphs are placed among priced services in the controlled study; retrieval is evaluated separately.

We introduce **Scrip**, a collective solver organized around this shared-budget rule. Each frozen agent can spend its account on a document, tool result, or peer rationale; the remaining balance supports its prediction. Verified training outcomes settle the account for the next task, while evaluation restores a frozen trained state.

The mechanism couples the two decisions directly. Buying evidence changes an agent's report and reduces the balance that supports its influence; verified outcomes update that balance across training tasks. In sealed mode, this is a cost-adjusted online ensemble and requires no backbone retraining. This coupling creates a measurable tradeoff: a specialist may need expensive evidence, and historical success can persist after expertise changes. Probability pooling combines relations represented in individual reports; agents who observe separate bits assign probability $1/2$ to their XOR until one shares the missing bit. Scrip treats rationale exchange as a priced service, allowing us to measure when communication adds information beyond probability pooling.

We specify the acquisition–prediction–settlement loop, derive the cost term in reliability weighting, vary scarcity, and test spending, error correlation, and non-oracle retrieval. The shared gain is $+0.8$ points at the original cap and $+3.6$ near 1.5 affordable evidence units (Figure 1).

The contribution is a decomposition of the measured gains. Selective acquisition and outcome-based weighting account for most of the improvement over a judge; sharing their budget adds a smaller, scarcity-dependent effect. Matched-report replay identifies the sealed system with a cost-adjusted pool, while a convex-hull bound identifies when signal-bearing communication is indispensable. Comparisons with deliberation, calibrated aggregation, changing expertise, and non-oracle retrieval locate the conditions under which the shared-budget rule helps.

2 Collective Inference with Shared Budgets

Scrip has 2 coupled processes. *Within a task*, agents acquire evidence and form a collective prediction. *Across training tasks*, verified outcomes update the accounts that govern the next acquisition and prediction. We begin with the setting and the simplest implementation, then explain why a sequential alternative requires a different contract. Section A gives the complete executable ordering.

2.1 The setting: partial evidence, a common question, and a resource limit

Agent i has a persona ρ_i, a frozen model θ, and beginning-of-task wealth $W_i > 0$. The default population contains $N = 16$ fixed identities. Agents share the task x and public context but may start with different private evidence shards. A posted-price menu offers additional document windows, tool results, or cached peer rationales. All agents can access the same menu subject to affordability, while differences in initial private observations create information asymmetry.

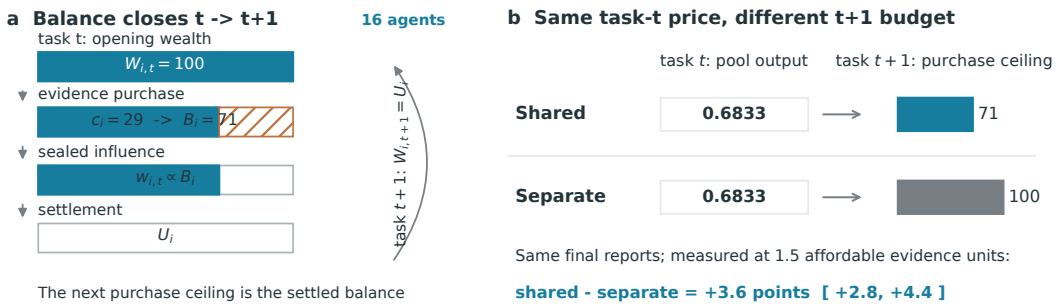

Figure 2: **The same balance pays for evidence and prediction influence.** In sealed mode, a purchase changes the report immediately and lowers the balance that weights that report. Settlement carries the resulting balance into the next training task; the separate-budget control keeps the prediction fund unchanged after the same purchase.

The output is a distribution $p \in \Delta_K$ over a common set of candidate answers and a decision $\arg\max_k p(k)$. Here Δ_K is the probability simplex and K is the number of candidates. Multiple-choice and binary forecasting tasks supply this set. For open answers, agents first propose answers; a deterministic canonicalizer merges equivalent strings, retains at most five distinct candidates using a fixed public hash, and appends OTHER, denoted $\perp$. Probability elicitation begins only after all agents receive the same ordered candidate list. Hash-based truncation avoids giving repeated nominations extra priority.

Two resource quantities have different roles. *Wealth* is an internal credit balance used to enforce access and influence. *Compute cost* records actual input and output tokens, reasoning tokens, tool charges, and latency. Every method receives the same global compute cap and evidence menu. An agent cannot exceed either its financial permission or the global cap, and an internal credit transfer does not erase a real model call.

2.2 Acquisition: Spending to Change a Prediction

Given its current evidence, agent i reports $q_i \in \Delta_K$. The acquisition controller observes this belief, purchased source IDs, the remaining balance, and service metadata, then chooses one service, a permitted bundle, or STOP. Content is delivered after payment. The agent revises its report and may return to the menu, up to the decision limit. Bundle purchases allow the controller to seek complementary evidence whose individual pieces may be uninformative. A peer rationale is one purchasable service: its generation and every buyer's reading are billed.

Let c_i be cumulative expenditure and $B_i = W_i - c_i > 0$ the remaining balance. A purchase changes two things at once: the report q_i, through the newly available information, and the balance B_i, through its cost. This is the defining coupling. The controller is a prompted policy, not an assumed Bayesian optimizer; its decisions are evaluated against random, round-robin, and lookahead controls. To keep all contracts well defined, validated reports are normalized and smoothed as $q_i \leftarrow (1 - K\epsilon)q_i + \epsilon\mathbf{1}$, with $\epsilon = 10^{-4}$.

The separate-budget control uses the same evidence menu and compute allowance, while purchases leave prediction weights unchanged. The two arms therefore differ in the opportunity cost that purchases impose on influence (Section C).

2.3 Prediction and Feedback: The Sealed Implementation

The sealed mode collects all final reports before revealing a collective price. A controller commits the same fraction $\kappa \in (0, 1]$ of each remaining balance, normally $\kappa = 0.5$. Writing s_i for the stake and $\widetilde{w}_i$ for the post-purchase share gives

$$s_i = \kappa B_i, \qquad \widetilde{w}_i = \frac{B_i}{\sum_j B_j}, \qquad p = \sum_i \widetilde{w}_i q_i. \tag{1}$$

The common fraction cancels from the price: the output is a linear probability pool weighted by remaining balances. Thus the stake is not a second, voluntary confidence report. At fixed q_i, an expensive purchase reduces its buyer's weight relative to an otherwise identical agent who did not purchase.

When the training contract resolves to outcome y, the pool pays each agent in proportion to its stake on y. Before any rescaling or optional population update, its balance is

$$U_i = B_i - s_i + s_i \frac{q_i(y)}{p(y)}. \tag{2}$$

The first two terms retain the uncommitted balance; the last is the winning payout. Predicting the outcome more strongly than the collective price increases the stake's return. Since the winning payouts sum to the total stake, settlement redistributes the post-purchase capital without creating it. The worked episode in Section E shows the corresponding accounting: evidence changes reports immediately, while its debit also changes the weight carried into settlement and the next task.

2.4 Price feedback: a collateralized sequential alternative

Sealed pooling gives a direct comparison with online ensembles. Scrip also supports a logarithmic market scoring rule (LMSR) (Hanson, 2003; 2007), which lets agents take turns observing and updating a public probability vector. Prices can reveal disagreement and prompt further purchases, and can also induce anchoring. Both modes are evaluated on every task, separating the price mechanism from the shared-budget factor.

The sequential contract starts at $p^0 = \mathbf{1}/K$. Each move adds the log-score contract and reserves its worst-case loss; the controller chooses the constrained log-score optimum under the collateral left after purchases. The complete feasibility rule and closed-form update are in Sections B and C.

At most two recorded rounds are allowed, with early stopping. Reserved losses cannot be spent again on evidence, and settlement is $U_i = W_i - c_i + z_i(y)$ with a separate maker reserve of $b \log K$. Both modes are evaluated on every task; changing the price mechanism is a separate factor from sharing the acquisition budget.

2.5 Across-task adaptation and frozen evaluation

The feedback loop is adaptive when settled training balances determine the next task's permissions. Core experiments keep personas fixed and set rent and renewal to zero. A common rescaling restores total wealth while preserving relative shares; the ledger and optional population changes are audited in Section C. Verified labels update accounts without backbone gradients, and learned baselines receive the same labels and tuning allowance.

Evaluation restores an independent frozen population snapshot for every item. Purchases and prediction remain active within the task, but test outcomes never update persistent wealth, personas, or selection; deployment-mode choices are locked on validation NLL. Candidate reopening, uncertainty diagnostics, and the complete accounting ledger are reported in the appendix.

3 Understanding the Coupling

The mechanism connects expenditure and influence by construction. We recover the classical no-purchase update, expose the extra acquisition term, and identify which information structures probability weighting can aggregate and which require communication. Proofs and counterexamples appear in Section B.

3.1 From reliability weighting to cost-adjusted influence

Without purchases, an account changes only through the quality of its reported probabilities relative to the pool. The following result explains why wealth can act as a reliability weight under a restricted, fixed-population protocol.

Proposition 1 (No-purchase expert comparison)**.** *Consider sealed pooling with a fixed population, positive reports, common constant κ, and no purchases, rent, rebates, replacements, capital injections, or selective participation. Let $w_i^t = W_i^t / \sum_j W_j^t$. Then*

$$w_i^{t+1} = w_i^t \left(1 - \kappa + \kappa q_i^t(y_t)/p_t(y_t)\right), \tag{3}$$

$$\mathcal{R}_i(T) := \sum_{t=1}^{T} \log \frac{q_i^t(y_t)}{p_t(y_t)} \leq \frac{\log(1/w_i^1)}{\kappa}. \tag{4}$$

At $\kappa = 1$, $w_i^{T+1} = w_i^1 \exp[\mathcal{R}_i(T)]$. In particular, excess average agent log-loss $\delta_i > 0$ gives $w_i^{T+1} = w_i^1 e^{-T\delta_i}$.

The bound concerns online wealth shares under the stated protocol. At $\kappa < 1$, wealth growth need not rank benchmark accuracy, calibration, or learned-pool performance.

Purchasing evidence breaks that simple identification: two equally accurate agents can incur different costs. Let a_i^t denote the fraction of agent i's wealth spent on task t, and ϕ_t the population's wealth-weighted average expenditure fraction. The next proposition isolates precisely where this difference enters.

Proposition 2 (Acquisition changes the comparator)**.** *Allow purchases $0 \leq c_i^t < W_i^t$, but keep the other restrictions above. Write $a_i^t = c_i^t / W_i^t$ and $\phi_t = \sum_i w_i^t a_i^t$. Then*

$$w_i^{t+1} = w_i^t \frac{1 - a_i^t}{1 - \phi_t} \left(1 - \kappa + \kappa \frac{q_i^t(y_t)}{p_t(y_t)}\right), \tag{5}$$

$$\mathcal{R}_i(T) \leq \frac{1}{\kappa} \left[\log \frac{1}{w_i^1} + \sum_t \log \frac{1 - \phi_t}{1 - a_i^t}\right]. \tag{6}$$

A common multiplicative rebate after settlement preserves these shares and the individual cost term.

The update has two factors: $(1 - a_i^t)/(1 - \phi_t)$ adjusts for expenditure relative to the population, and the remaining factor adjusts for predictive performance. Spending a larger fraction than the population average reduces influence before forecast quality is rewarded, and common rescaling preserves that relative penalty. This cost term motivates the separate-budget comparison; whether its effect improves the collective answer depends on the evidence and agent behavior. On identical reports and expenditures, a cost-adjusted fractional online pool is exactly Scrip sealed.

The cost factor predicts where an effect can emerge. When purchases are cheap relative to wealth, a_i^t and ϕ_t are small and $\log[(1 - a_i^t)/(1 - \phi_t)] \approx \phi_t - a_i^t$, so the two arms nearly coincide. When useful evidence is unaffordable, neither arm can act on it. Between these limits, purchases remain feasible and their debit changes the influence available afterward. Evidence quality and agent behavior determine the sign and location of the accuracy peak, which we measure experimentally.

3.2 What must a purchase accomplish?

A report can change without becoming more useful. The full-Kelly reference in Proposition 3 formalizes the tradeoff: evidence is worthwhile when its conditional information exceeds the log opportunity cost of its fee. We compare the prompted controller with this price-taking benchmark using forced buy/no-buy branches and bundle controls.

3.3 When must agents exchange more than probabilities?

Every sealed output lies in the convex hull of its reports: $p(y) \leq \max_i q_i(y)$. The pool can outperform each fixed agent across tasks, while its confidence on any one outcome remains bounded by the strongest report on that outcome.

For independent uniform bits A, B and $Y = A \operatorname{XOR} B$, seeing either bit alone leaves the answer posterior at $1/2$. Truthful probability-only pooling, or myopic LMSR trading from price $1/2$, therefore reveals no answer information. Sharing the missing bit permits exact reconstruction. General strategic real-valued reports can encode the bits, which constitutes a richer communication protocol. This boundary motivates both paid rationales and the exact information experiments in Section D.10.

The sequential contract's telescoping payout bounds maker liability by $b \log K$ from a uniform initial price. Using trading gains to fund information costs additionally requires behavioral assumptions (Section B).

These results lead to three distinct tests. The cost term calls for shared-versus-separate budgets; the acquisition tradeoff calls for matched purchase policies and forced branches; the communication boundary calls for controlled information structures. We use this progression to organize the evaluation.

4 Experiments: Locating the Value of a Shared Budget

The evaluation follows the acquisition–prediction loop from opportunity to outcome. We locate the gap between a judge and systems that can buy and weight evidence, vary scarcity, then test spending, dependence, retrieval, replay, and information-structure boundaries. Complete matrices are in Section D.

4.1 Setup: matching access, adaptation, and computation

Tasks and partiality. We consider Qwen3-8B and GPT-5-mini on MATH-500, AIME 2024–2025, GPQA-Diamond, controlled 2Wiki, and binary forecasting. The first three limit gains from new evidence; the last two emphasize unevenly distributed documents. The expanded Qwen studies use 1,000 2Wiki questions. In the controlled task, gold-supporting paragraphs are placed among priced services to isolate allocation; non-oracle retrieval is reported separately (Section 4.4).

Compared mechanisms. Single-CoT, voting, confidence pooling, Debate, MoA, and Judge provide familiar reference points. The decisive controls are stronger: Bayesian and fractional online pools test learned reliability; calibrated logit stacking tests learned probability aggregation; separate budgets test the purchase debit; and InfoDelphi tests deliberation with public/private evidence routing and rationale exchange. LMAD, DEAR, Consilience, and two internal-market controls extend the comparison. Every stronger control can use the same evidence service menu and acquisition interface. Published methods are adapted to this protocol and evaluated on our task suite.

Adaptation and measurement. The core uses fixed personas, five independent training streams of 300 resolved draws, and three decoding seeds per original test item. Learned controls receive the same resolved labels and validation allowance. The $1\times$ token caps are fixed in advance: 64K for MATH/GPQA, 80K for 2Wiki/forecasting, and 96K for AIME. Controller calls, rationale generation and reading, hidden branches, retries, tools, and reopening are billed. Training costs are reported separately. GPQA uses a main-minus-Diamond training/validation split; forecasting uses event-cluster separation and as-of snapshots (Section D).

Comparisons and uncertainty. Both sealed and LMSR are evaluated on every original task. Validation-locked selection uses sealed for mathematics/science and LMSR for evidence tasks. We average decoding seeds within each item and apply crossed stream/item resampling to paired differences; AIME still has 60 independent question units. The original primary comparisons are shared versus separate budgets on 2Wiki NLL and forecasting Brier. The scarcity scan reports paired accuracy at five affordability levels. Proper scores, calibration, and cost remain separate checks of prediction quality.

4.2 A larger gain from acquisition than from coupling

Figure 1a begins with four controlled Qwen 2Wiki configurations: Judge at 57.0%, separate budgets at 64.0%, shared sealed budgets at 64.8%, and shared sequential trading at 65.4%. The seven-point difference from Judge combines acquisition with outcome-based weighting. Selective acquisition with uniform pooling reaches 62.4%, while frozen learned weights without adaptive access reach 63.4%. Both capabilities matter before the shared debit is introduced; the complete ablation is in the appendix.

On the same controlled 2Wiki task with GPT-5-mini, selected Scrip reaches 77.9% versus 76.5% for separate budgets. Judge remains strongest on MATH, while Scrip and InfoDelphi trade small leads

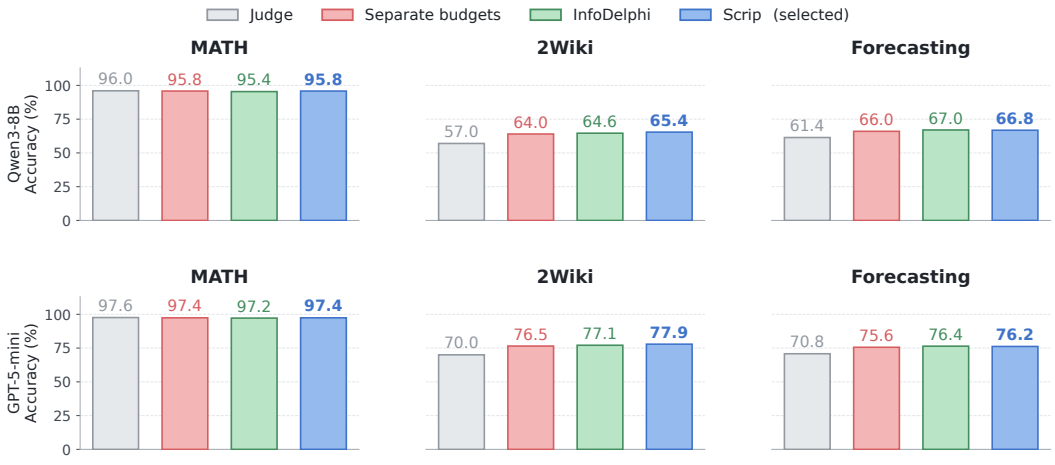

Figure 3: **Selected cross-task results.** Bars show the locked configurations for the Qwen3-8B and GPT-5-mini tracks on mathematics, controlled 2Wiki, and forecasting. The full method matrix, uncertainty intervals, and additional tasks are reported in Appendix D.

across the remaining tasks. Appendix D reports the full matrix and intervals; the comparison here isolates when debiting evidence purchases from influence helps a capable ensemble.

4.3 Scarcity exposes the purchase debit

Holding sealed aggregation fixed, the original Qwen shared-versus-separate comparison is 64.8% versus 64.0% ($+0.80$ points, interval $[+0.25, +1.35]$). Its small default effect motivates varying the available evidence budget.

Let k denote the approximate number of evidence units affordable per task. Figure 1b shows a non-monotonic response across five levels: with $k \geq 6$, sharing gains $+0.8$ points; near $k \approx 1.5$, separate budgets fall to 58.4% and shared budgets to 62.0%, a $+3.6$-point paired difference $[+2.8, +4.4]$; when $k \approx 0.4$, the difference contracts to $+1.3$. Absolute accuracy falls as access tightens in both arms. The useful regime is therefore a budget that forces selection among available evidence, yet still permits consequential purchases. The underlying five-point measurements remain available in the appendix.

The spending diagnostic helps interpret the peak. When gold-supporting passages are priced above baseline and decoys below it, the shared arm reaches 63.4% versus 60.2% for separate budgets, with a paired $+3.2$-point interval $[+2.4, +4.0]$. The shared arm allocates 57% of its evidence budget to gold-supporting passages, compared with 41% for separate budgets, and 21% to decoys, compared with 33%. Buyers see service metadata, not these gold labels. At the same $k \approx 1.5$ setting, a spend-matched relevance router reaches 60.8% against 62.0% for shared budgets. The advantage persists against designer-directed allocation at matched evidence expenditure.

Error dependence moves with access. Mean pairwise correctness correlation is 0.63 with homogeneous input, 0.49 with separate budgets, and 0.41 with shared budgets. Using $N_{\text{eff}} = N/[1+(N-1)\rho]$ for $N = 16$, these correspond to 1.53, 1.92, and 2.24 exchangeable-equivalent independent agents. This descriptive contrast is consistent with differentiated evidence paths. Since report quality and item difficulty also vary, the comparison describes how error correlation co-varies with access; it does not identify the causal effect of a purchase. It shows why adding identical agents alone leaves the initial pooling gap unresolved (Begin et al., 2026). Price feedback is a secondary channel: LMSR raises the original controlled evidence score from 64.8% to 65.4% on Qwen, while renewal reaches 65.8% at extra training cost. Fixed personas therefore define the core system; supplying all evidence tiers costs $1.42\times$ but gives only 64.1%.

4.4 What survives ordinary retrieval?

The controlled construction measures allocation when key passages are available for purchase. In the original non-oracle evaluation, Scrip LMSR scores 60.4%, close to InfoDelphi's 60.0%, down from 65.4% under controlled access. A separate expanded retrieval evaluation yields 62.6% for the shared account, versus 60.9% for separate budgets and 60.6% for a spend-matched relevance router. The paired difference from the strongest listed comparator is $+1.7$ points $[+0.8, +2.6]$. Retrieval noise narrows the advantage relative to the peak controlled-scarcity result. The two non-oracle protocols are summarized in Appendix D; the 60.4% and 62.6% results come from separate evaluations.

4.5 The sealed rule reduces to online aggregation

The sealed formula in Equations (1) and (5) yields an exact replay test. On a matched trace fixing candidates, purchases, fees, reports, and labels, Scrip sealed and a cost-adjusted fractional pool both score 64.8% and make identical predictions. The shared account changes future permissions and the reports produced by a closed loop; conditional on the same trace, it is an ordinary online aggregation recurrence. Learned weighting is therefore compared on the same acquired information, isolating its aggregation operator.

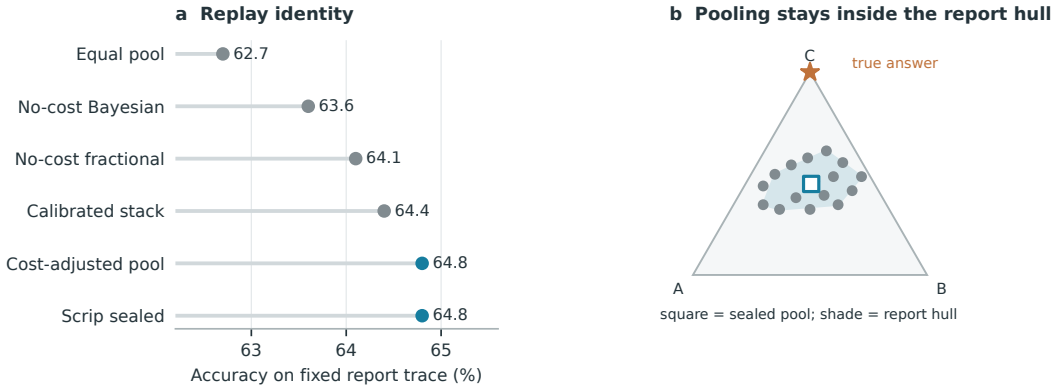

Figure 4: **Replay identifies the sealed operator, and pooling stays inside the report hull. (a)** Matched Qwen replay gives identical 64.8 points for sealed Scrip and the cost-adjusted pool. **(b)** A report set in a three-outcome simplex illustrates $p(y) \leq \max_i q_i(y)$ when the true answer lies outside the report hull.

Additional diagnostics. The probability, resource, population, and information-structure diagnostics are reported in Section D. They show how calibration and cost affect interpretation, how historical weighting reacts to changing expertise, and where identical evidence or missing complementary signals limit the system. The main evidence chain therefore ends with the measured replay identity and the convex-hull boundary; the full stress-test record remains available without interrupting the core result.

5 Related Work

Scrip connects economic coordination, outcome-based aggregation, and selective information acquisition. The key distinction is what the resource buys, who assigns evidence, and whether its expenditure changes predictive influence.

Economic coordination and internal markets. Galanis (2026) study private-signal aggregation by LLM traders in LMSR markets; Begin et al. (2026) show how persistent wealth and correlated errors constrain a question-answering crowd; and Qian (2026) analyze wealth concentration and XOR information limits. Scrip adds a priced choice of *which information to acquire* before reporting, then measures the downstream effect when the same account pays for that choice. Economy of

Minds (Qi et al., 2026) supplies the closest population template (frozen backbones, personas, wealth, amendment, mutation, and frozen evaluation), but money there allocates actions whereas here it buys evidence and determines pooling weight. PolySwarm (Barot & Borkhatariya, 2026) combines confidence with external market prices rather than endogenous evidence acquisition.

Deliberation and information structure. Debate (Du et al., 2024; Liang et al., 2024; Khan et al., 2024), role-based agents (Li et al., 2023; Hong et al., 2023; Wu et al., 2023), self-consistency (Wang et al., 2023), and MoA (Wang et al., 2024a) are foundational controls. LMAD (Gao et al., 2026), DEAR (Wu et al., 2026), Meta-Moderator (Hu et al., 2026), and Consilience (Babu et al., 2026) regulate communication. InfoDelphi (Li et al., 2026) is the closest allocation control: it routes public/private evidence and exchanges rationales, while our spend-matched variant tests whether endogenous purchases match that assignment at equal cost. Natural-language mechanisms (Della Penna, 2024) motivate the complementary-signal tests.

Online learning and probability aggregation. Bayesian averaging (Hoeting et al., 1999), stacking (Wolpert, 1992), Hedge (Freund & Schapire, 1997; Cesa-Bianchi & Lugosi, 2006), and Kelly markets (Kelly, 1956; Beygelzimer et al., 2012; Storkey, 2011; Hu & Storkey, 2014) already learn from outcomes. These connections motivate our matched-report replay and learned-pool controls; proper scoring and calibration remain distinct from weight updates (Gneiting & Ranjan, 2011; Gneiting & Raftery, 2007).

Acquisition, rationality, and evaluation. Costly information underlies metareasoning, informational substitutes and complements, adaptive submodularity, and budgeted learning (Grossman & Stiglitz, 1980; Russell & Wefald, 1991; Chen & Waggoner, 2017; Golovin & Krause, 2011; Fainman et al., 2019; Efroni et al., 2021). Information bazaars and evidence markets consider buying or rewarding information (Rahaman et al., 2024; Sudhir & Tran-Thanh, 2026; Hossain et al., 2026); Scrip fixes seller prices to isolate the buyer's coupling rule and tests purchase behavior directly (Pal et al., 2025). Hindcast and forecasting pipelines guide temporal hygiene (Ye et al., 2026; Halawi et al., 2024); further context appears in Section F.

6 Discussion and Conclusion

Collective reasoning under a resource limit begins before aggregation: a system must decide which evidence its agents should acquire. Scrip makes that choice and the ensuing allocation of influence jointly observable. In the controlled Qwen 2Wiki sequence, the move from Judge to an acquiring, outcome-weighted separate-budget system is seven points, while the shared sealed account adds 0.8 points at the original budget. The latter difference rises to 3.6 points when a task affords roughly 1.5 evidence units, then shrinks when almost no evidence can be bought. The debit is consequential in this intermediate-scarcity regime.

The evidence supports a conditional design rule: make evidence expenditure compete with influence when agents can choose among useful and redundant sources and the budget makes that choice matter. Shared budgets spend more on gold-supporting passages under heterogeneous prices and reduce correctness correlation; in a separate non-oracle retrieval evaluation, the advantage over the strongest listed control is 1.7 points. Replay identifies sealed Scrip with a cost-adjusted fractional pool on identical reports, while the convex-hull and XOR tests locate tasks where complementary signals require communication. Selective deliberation and calibrated pooling remain strong on forecasting, Judge leads on mathematics, and resource savings should be assessed alongside adaptation costs.

The mechanism uses verifiable outcomes, a specified evidence interface, and auditable costs. Adaptive sellers lower accuracy in both budget arms; collusion, unreliable labels, changing expertise, and strategic evidence contents remain open problems. For consequential use, Scrip needs paid rationale exchange, provenance-aware acquisition, task-dependent forgetting, calibration, and human review. Appendix G details the limitations and failure cases.

AI Use Statement

Generative AI tools assisted literature discovery, language editing, mathematical checks, visualization code, and LaTeX compilation. The authors remain responsible for verification.

REFERENCES

Isaiah Andrews and Suproteem Sarkar. Dutch books for language models. *arXiv preprint arXiv:2609.02797*, 2026.

Kenneth J. Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D. Nelson, George R. Neumann, Marco Ottaviani, Thomas C. Schelling, Robert J. Shiller, Vernon L. Smith, Erik Snowberg, Cass R. Sunstein, Paul C. Tetlock, Philip E. Tetlock, Hal R. Varian, Justin Wolfers, and Eric Zitzewitz. The promise of prediction markets. *Science*, 320(5878):877–878, 2008.

Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, 63(3):691–706, 2017.

Abhijith Babu, Ramneet Kaur, Vishal Pramanik, Olivera Kotevska, Nathaniel D. Bastian, Susmit Jha, Sunny Raj, Yanzhao Wu, Sumit Kumar Jha, and Anirban Roy. Consilience: Conformally calibrated communication control for hidden-profile multi-agent reasoning. *arXiv preprint arXiv:2608.20564*, 2026.

Marc Bara. Epistemic sybil resistance: Multiplying ai agents without multiplying evidence. *arXiv preprint arXiv:2609.01873*, 2026.

Rajat M. Barot and Arjun S. Borkhatariya. PolySwarm: A multi-agent large language model framework for prediction market trading and latency arbitrage, 2026. arXiv:2604.03888.

Eric B. Baum. Toward a model of intelligence as an economy of agents. *Machine Learning*, 35(2): 155–185, 1999.

Eric B. Baum and Igor Durdanovic. Evolution of cooperative problem solving in an artificial economy. *Neural Computation*, 12(12):2743–2775, 2000.

James Begin, Brendan Gho, Suman Muppavarapu, Tyson Tsay, Atharva Mohan, Afnan Shaik, Ruizhe Li, Vasu Sharma, and Archana Vaidheeswaran. Preference optimization drives monoculture in llm prediction markets, 2026. URL https://arxiv.org/abs/2606.26583. arXiv:2606.26583.

Joyce Berg, Robert Forsythe, Forrest Nelson, and Thomas Rietz. Results from a dozen years of election futures markets research. In *Handbook of Experimental Economics Results*, volume 1, pp. 742–751. Elsevier, 2008.

Alina Beygelzimer, John Langford, and David M. Pennock. Learning performance of prediction markets with Kelly bettors. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1317–1318, 2012.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. Scalable AI safety via doubly-efficient debate, 2023. arXiv:2311.14125.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail?, 2025. arXiv:2503.13657.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. Put your money where your mouth is: Evaluating strategic planning and execution of LLM agents in an auction arena, 2023a. arXiv:2310.05746.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance, 2023b. arXiv:2305.05176.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors, 2023c. arXiv:2308.10848.

Yiling Chen and Bo Waggoner. Informational substitutes, 2017. URL https://arxiv.org/abs/1703.08636. arXiv:1703.08636.

Nicolas Della Penna. Natural language mechanisms via self-resolution with foundation models, 2024. URL https://arxiv.org/abs/2407.07845. arXiv:2407.07845.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference (WWW)*, 2024.

Yonathan Efroni, Nadav Merlis, Aadirupa Saha, and Shie Mannor. Confidence-budget matching for sequential budgeted learning, 2021. URL https://arxiv.org/abs/2102.03400. arXiv:2102.03400.

Eran Fainman, Bracha Shapira, Lior Rokach, and Yisroel Mirsky. Online budgeted learning for classifier induction, 2019. URL https://arxiv.org/abs/1903.05382. arXiv:1903.05382.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Spyros Galanis. Information aggregation with ai agents, 2026. URL https://arxiv.org/abs/2604.20050. arXiv:2604.20050.

Weijun Gao, Xiang Ding, Haoyang Liu, and Tiancheng Xing. Where reasoning diverges: Localized multi-agent debate for multi-hop question answering. *arXiv preprint arXiv:2608.01463*, 2026.

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

Tilmann Gneiting and Roopesh Ranjan. Combining predictive distributions, 2011. URL https://arxiv.org/abs/1106.1638. arXiv:1106.1638.

Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011. arXiv:1003.3967.

Sanford J. Grossman and Joseph E. Stiglitz. On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3):393–408, 1980.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models, 2024. arXiv:2402.18563.

Robin Hanson. Combinatorial information market design. *Information Systems Frontiers*, 5(1): 107–119, 2003.

Robin Hanson. Logarithmic market scoring rules for modular combinatorial information aggregation. *The Journal of Prediction Markets*, 1(1):3–15, 2007.

Friedrich A. Hayek. The use of knowledge in society. *The American Economic Review*, 35(4): 519–530, 1945.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

John H. Holland. Properties of the bucket brigade algorithm. In *Proceedings of the 1st International Conference on Genetic Algorithms*, pp. 1–7, 1985.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework, 2023. arXiv:2308.00352.

Safwan Hossain, Gabriel Andrade, Chengqi Zang, and Yiling Chen. Evidence markets, 2026. arXiv:2606.07434.

Jinli Hu and Amos Storkey. Multi-period trading prediction markets with connections to machine learning. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pp. 1773–1781, 2014.

Wentao Hu, Zhuoyue Wan, Jinhao Shen, Chen Jason Zhang, Xiaoyong Wei, and Qing Li. Meta-moderator: Empowering multi-agent debate with meta-cognition, 2026. arXiv:2608.23029.

Evan Hubinger. AI safety via market making. AI Alignment Forum, 2020. URL https://www.alignmentforum.org/posts/YWwzccGbcHMJMpT45/ai-safety-via-market-making.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate, 2018. arXiv:1805.00899.

Athul Paul Jacob, Yikang Shen, Gabriele Farina, and Jacob Andreas. The consensus game: Language model generation via equilibrium search. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

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, 2022. arXiv:2207.05221.

John L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35(4): 917–926, 1956.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Tarun Kota. Design and evaluation of multi-agent ai oracle systems for prediction market resolution, 2026. URL https://arxiv.org/abs/2605.30802. arXiv:2605.30802.

Nicolas S. Lambert, John Langford, Jennifer Wortman, Yiling Chen, Daniel Reeves, Yoav Shoham, and David M. Pennock. Self-financed wagering mechanisms for forecasting. In *Proceedings of the 9th ACM Conference on Electronic Commerce (EC)*, pp. 170–179, 2008.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Yuante Li, Yicheng Tao, Kate Zhang, Taozhi Wang, Gefei Gu, and Yaxin Zhou. Diverse evidence, better forecasts: Multi-agent deliberation under information asymmetry, 2026. URL https://arxiv.org/abs/2607.01661. arXiv:2607.01661.

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

Yaokun Liu, Yifan Liu, Daniel Yue Zhang, Ruichen Yao, Zelin Li, and Dong Wang. PropUQ-MAS: Propagation-aware uncertainty quantification for llm multi-agent systems, 2026. arXiv:2608.22130.

Barbara Mellers, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E. Scott, Don Moore, Pavel Atanasov, Samuel A. Swift, Terry Murray, Eric Stone, and Philip E. Tetlock. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5):1106–1115, 2014.

Mark S. Miller and K. Eric Drexler. Markets and computation: Agoric open systems. In Bernardo A. Huberman (ed.), *The Ecology of Computation*, pp. 133–176. North-Holland, 1988.

Marvin Minsky. *The Society of Mind*. Simon and Schuster, 1988.

Michael Ostrovsky. Information aggregation in dynamic markets with strategic traders. *Econometrica*, 80(6):2595–2647, 2012.

Arka Pal, Teo Kitanovski, Arthur Liang, Akilesh Potti, and Micah Goldblum. Knowing what you know is not enough: Large language model confidences don't align with their actions, 2025. URL https://arxiv.org/abs/2511.13240. arXiv:2511.13240.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

Zhenting Qi, Huangyuan Su, Ao Qu, Chenyu Wang, Yu Yao, Han Zheng, Kushal Chattopadhyay, Guowei Xu, Zihan Wang, Weirui Ye, Vijay Janapa Reddi, Ju Li, Paul Pu Liang, Himabindu Lakkaraju, Sham Kakade, and Yilun Du. Economy of minds: Emerging multi-agent intelligence with economic interactions, 2026. arXiv:2606.02859.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Cheng Qian. Information limits and attractor dynamics in economies of frontier llm agents: A preregistered test, 2026. URL https://arxiv.org/abs/2607.06001. arXiv:2607.06001.

Nasim Rahaman, Martin Weiss, Manuel Wüthrich, Yoshua Bengio, Li Erran Li, Chris Pal, and Bernhard Schölkopf. Language models can reduce asymmetry in information markets, 2024. arXiv:2403.14443.

Itzhak Rasooly and Roberto Rozzi. How manipulable are prediction markets?, 2025. URL https://arxiv.org/abs/2503.03312. arXiv:2503.03312.

Stuart Russell and Eric Wefald. Principles of metareasoning. *Artificial Intelligence*, 49(1–3):361–395, 1991.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Jürgen Schmidhuber. The neural bucket brigade: A local learning algorithm for dynamic feedforward and recurrent networks. *Connection Science*, 1(4):403–412, 1989.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, et al. Towards understanding sycophancy in language models, 2023. arXiv:2310.13548.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters, 2024. arXiv:2408.03314.

Amos Storkey. Machine learning markets. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 716–724, 2011.

Abhimanyu Pallavi Sudhir and Long Tran-Thanh. Extrapolating volition with recursive information markets, 2026. arXiv:2604.08606.

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 (EMNLP)*, 2023.

Michael Todasco. Going all-in on LLM accuracy: Fake prediction markets, real confidence signals, 2025. arXiv:2512.05998.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities, 2024a. arXiv:2406.04692.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the answer? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024b.

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 *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

Michael P. Wellman. A market-oriented programming environment and its application to distributed multicommodity flow problems. *Journal of Artificial Intelligence Research*, 1:1–23, 1993.

Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2): 107–126, 2004.

Justin Wolfers and Eric Zitzewitz. Interpreting prediction market prices as probabilities. Working Paper 12200, National Bureau of Economic Research, 2006.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

Hao Wu, Shoucheng Song, Chang Yao, Haoyu Wang, Huaiyu Wan, Youfang Lin, and Kai Lv. Group perspective matters: Regulating debate relationships can mitigate blind conformity in multi-agent debate. *arXiv preprint arXiv:2608.03648*, 2026.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation, 2023. arXiv:2308.08155.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 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 *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

Xiao Ye, Jacob Dineen, Evan Zhu, Shijie Lu, Kevin Song, and Ben Zhou. Hindcast: Replaying prediction markets to evaluate llm forecasters, 2026. URL https://arxiv.org/abs/2607.14051. arXiv:2607.14051.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

A EXECUTABLE PROTOCOL

The implementation order matters: candidate proposal precedes probability elicitation; the market branch determines the contract; reopening voids a contract before starting another; verification is unavailable to acquisition or prediction. Algorithms below specify these dependencies.

Algorithm 1 Episode with one optional reopening

Input: Task x, snapshot $(\rho_i, W_i)_{i=1}^N$, menu, mode, token cap
$c_i \leftarrow 0$; evidence cache $E_i \leftarrow$ permitted public/private context
for attempt $a = 1, 2$ **do**
  Propose, canonicalize, truncate distinct answers by public hash; append OTHER
  Reset contracts: $z_i \leftarrow 0$, $u_i \leftarrow 0$; elicit q_i over fixed candidates
  if mode = sealed **then**
    Run private acquisition controller while cap allows; deduct fees; revise reports
    $s_i \leftarrow \kappa(W_i - c_i)$; $p \leftarrow \sum_i s_i q_i / \sum_i s_i$
  else
    $p \leftarrow \mathbf{1}/K$
    for at most R rounds, agents in recorded random permutation **do**
      Agent i may buy $d < W_i - c_i$ only if $u_i \le \kappa(W_i - c_i - d)$; revise q_i
      $d_i \leftarrow \kappa(W_i - c_i) - u_i$; solve constrained report r_i
      $z_i \leftarrow z_i + b \log(r_i/p)$; $u_i \leftarrow u_i + \ell(p, r_i)$; $p \leftarrow r_i$
    end for
  end if
  if $\arg\max p \neq \perp$ **or** $a = 2$ **then**
    Compute $\bar{q}$, J, R; **return** prediction, candidates, fees, mode-specific contracts
  end if
  Void all trading contracts; retain evidence charges and cache; charge reopening calls
end for

Algorithm 2 Training with a complete financial ledger

Input: Disjoint resolved training stream, $N = 16$, $W_0 = 100$, optional renewal flag
Initialize personas and wealth; charge persona-generation compute
for each training task x_t **do**
  Run Episode and record prediction before revealing verified answer
  Map answer into the final candidate set, using OTHER only for contract settlement
  if mode = sealed **then**
    $U_i \leftarrow W_i - c_i - s_i + s_i q_i(y)/p(y)$; maker entry $M \leftarrow 0$
  else
    $U_i \leftarrow W_i - c_i + z_i(y)$; maker entry $M \leftarrow -\sum_i z_i(y)$
  end if
  Assert $\sum_i U_i + \sum_i c_i + M = \sum_i W_i$ and $U_i \ge 0$
  if renewal enabled **then**
    Deduct capped rent; retire eligible balances to house; create new persona IDs
    Credit logged entrant endowments; charge amendment, mutation, and generation calls
  end if
  Rescale all live balances by a common factor to total NW_0; log net house transfer
  Save persona versions, wealth, label exposure, costs, and random-order seeds
end for

Algorithm 3 Evaluation and paired diagnostic replay

Input: Frozen training state, validation-locked settings, test items, method-paired seeds
1: **for** each test item **do**
2: Restore independent snapshot; run Episode within the shared resource cap
3: Save candidate list, prediction, probabilities, J, R, evidence paths, and all usage
4: Apply preselected abstention rule; OTHER is never credited as an answer
5: Score only after output is committed; do not update persistent state
6: Optionally replay the identical report trace through alternate aggregators
7: **end for**
8: Average decoding seeds within item; resample streams and item/event clusters for intervals
9: For purchase diagnostics, branch frozen states into forced buy/no-buy; charge analysis separately

B Proofs, Conditions, and Counterexamples

The proofs follow the mechanism: settlement defines the weight update, purchases alter the comparator, and information structure determines what reports can convey. They establish accounting identities and conditional bounds; the experiments assess behavioral and calibration outcomes.

B.1 Sealed settlement and the cost correction

Proof of Proposition 1. Let $S = \kappa \sum_i W_i$. Since $p(y) = \sum_i W_i q_i(y) / \sum_i W_i$, the sum of winning payouts is $\sum_i \kappa W_i q_i(y)/p(y) = S$. Total wealth is conserved, giving Equation (3). For $x > 0$, concavity of the logarithm gives $\log(1 - \kappa + \kappa x) \geq \kappa \log x$. Therefore

$$\log \frac{w_i^{T+1}}{w_i^1} = \sum_{t=1}^{T} \log \left(1 - \kappa + \kappa \frac{q_i^t(y_t)}{p_t(y_t)}\right) \geq \kappa \mathcal{R}_i(T).$$

Since $w_i^{T+1} \leq 1$, rearranging proves the bound. At $\kappa = 1$ the logarithmic inequality is equality, which gives the identity with index $T + 1$. Neither independence between agents nor stochastic outcomes is required. Reports may depend on the past, provided the identity of the comparator and the update rule remain fixed. □

Proof of Proposition 2. After purchases the available total is $\sum_i B_i^t = (1 - \phi_t) \sum_i W_i^t$. Applying the previous payout identity to these balances conserves this *post-purchase* total. Normalizing U_i therefore gives Equation (5). Taking logarithms and telescoping yields

$$\log \frac{w_i^{T+1}}{w_i^1} = \sum_t \log \frac{1 - a_i^t}{1 - \phi_t} + \sum_t \log \left(1 - \kappa + \kappa \frac{q_i^t(y_t)}{p_t(y_t)}\right).$$

Use the same concavity inequality and $\log w_i^{T+1} \leq 0$. A common factor multiplying every U_i cancels from normalized shares, so a pro-rata multiplicative refund preserves this formula, including its cost term. Cancellation requires this proportional multiplicative form; additive refunds, rent, replacement, and unequal injections alter normalized shares. □

For example, $W_1 = W_2 = 100$, $c_1 = 90$, $c_2 = 0$, and identical reports give balances $(10, 100)$ before recycling, whatever the common stake fraction. Recycling the full 90 pro-rata gives $(200/11, 2000/11)$. Differential expenditure therefore changes relative wealth even when the total fee is returned. If a rebate or rescaling also changes acquisition behavior, unchanged current shares can still lead to different future trajectories.

B.2 Information value and the fractional expansion

The acquisition controller must decide whether a possible belief improvement compensates for its cost. The following price-taking reference separates that decision from finite-market feedback.

Proposition 3 (Scoped value of information)**.** *For a coherent Bayesian agent, let* $q_e(y) = \Pr(Y = y \mid E = e, x)$, $q_0 = \mathbb{E}_e q_e$, *and fix an exogenous strictly positive price* $\hat{p}$ *that is the same with and*

without evidence. For full-Kelly allocation, external nonrefundable cost $c < W$ improves expected log wealth exactly when

$$I(E; Y \mid x) > \log \frac{W}{W - c}. \tag{7}$$

For the mechanically enforced fractional allocation $f = q$, define

$$G_\kappa(q, \hat{p}) = \sum_y q(y) \log \left(1 - \kappa + \kappa \frac{q(y)}{\hat{p}(y)}\right). \tag{8}$$

The exact information uplift is $\Delta_\kappa = \mathbb{E}_e G_\kappa(q_e, \hat{p}) - G_\kappa(q_0, \hat{p})$, and the purchase criterion is $\Delta_\kappa > \log[W/(W - c)]$. As $\kappa \to 0$,

$$\Delta_\kappa = \kappa \sum_y \frac{\mathrm{Var}_e[q_e(y)]}{\hat{p}(y)} + O(\kappa^2). \tag{9}$$

Equality means indifference. This price-taking analysis provides a reference for evaluating prompted purchases in finite markets. Complementary evidence calls for bundle selection or lookahead. The fixed-prefix result applies to deterministic fixed-depth policies; adaptive stopping requires additional structure.

Proof of Proposition 3. At external odds $\hat{p}$, full allocation f and report q give expected log wealth $\log W + \sum_y q(y) \log[f(y)/\hat{p}(y)]$. Over $f \in \Delta_K$, the f-dependent term is maximized at $f = q$, because its gap to the maximum is $\mathrm{KL}(q \,\|\, f)$. Without purchase the optimum is $\log W + \mathrm{KL}(q_0 \,\|\, \hat{p})$; after purchase it is $\log(W - c) + \mathbb{E}_e \mathrm{KL}(q_e \,\|\, \hat{p})$. Coherence implies $\mathbb{E}_e q_e = q_0$, and cancellation of the price terms gives

$$\mathbb{E}_e \mathrm{KL}(q_e \,\|\, \hat{p}) - \mathrm{KL}(q_0 \,\|\, \hat{p}) = H(q_0) - \mathbb{E}_e H(q_e) = I(E; Y \mid x).$$

Subtracting the no-purchase baseline proves the strict full-Kelly criterion. For a fixed fractional rule, ending wealth is $W[1 - \kappa + \kappa q(Y)/\hat{p}(Y)]$. Subtracting the two expected logarithms gives the stated exact uplift and log-cost threshold. On finite K with $\min_y \hat{p}(y) > 0$, the ratios are uniformly bounded, so Taylor expansion and expectation can be interchanged near $\kappa = 0$:

$$G_\kappa(q, \hat{p}) = \kappa \left(\sum_y \frac{q(y)^2}{\hat{p}(y)} - 1 \right) + O(\kappa^2).$$

Subtracting the q_0 expression leaves $\kappa \sum_y \mathrm{Var}_e[q_e(y)]/\hat{p}(y) + O(\kappa^2)$. □

For $\hat{p} = q_0 = (1/2, 1/2)$ and equally likely posteriors $(0.9, 0.1)$ and $(0.1, 0.9)$, $I = 0.368064$, whereas $\Delta_\kappa/\kappa \to 0.64$. At $\kappa = 0.5$, the exact uplift is 0.251742 and $\kappa I = 0.184032$. The optimizer $f = q$ characterizes full-Kelly allocation under external odds; at fractional stake or endogenous parimutuel odds, $f = q$ is the enforced policy analyzed here.

Lemma 1 (Fixed-prefix stopping only)**.** *Let the purchase decision choose a deterministic prefix of depth ℓ before observing any evidence. Let $C_\ell = \sum_{j \le \ell} c_j < W$ and $I_\ell = I(E_{1:\ell}; Y \mid x)$. Suppose increments $I_\ell - I_{\ell-1}$ are nonincreasing and log-cost increments $\log[(W - C_{\ell-1})/(W - C_\ell)]$ are nondecreasing. Then a maximizer of $I_\ell + \log(1 - C_\ell/W)$ is obtained by including all strictly positive net increments, with arbitrary treatment of zero increments.*

Proof. The difference between consecutive objective values is the information increment minus the log-cost increment. Under the hypotheses this difference is nonincreasing, so the objective increases until its difference first becomes nonpositive and never increases afterward. The result applies to deterministic fixed-depth prefixes. □

For adaptive selection, a history h has Bellman value $V(h, B) = \max\{U(h, B), \max_{e: c_e < B} \mathbb{E}[V(h \cup \{e\}, B - c_e) \mid h]\}$, with U the stop utility. A one-step policy can miss complementary evidence: in XOR, each first marginal information gain is zero but the two-bit bundle supplies $\log 2$. Adaptive submodularity supplies one structural condition for greedy guarantees (Golovin & Krause, 2011).

B.3 LMSR IMPLEMENTATION AND CONDITIONAL EXPENDITURE BOUND

Lemma 2 (Feasible reports, solvency, and maker loss)**.** *For the implementation in Section 2.4, the feasible set is nonempty, the constrained report is unique, $z_i(k) \geq -u_i$ for every outcome, and $U_i \geq (1-\kappa)(W_i - c_i) \geq 0$. Aggregate maker payout on a settled contract is at most $b \log K$.*

Proof. The current p is feasible since $p(k) \geq \epsilon$ and $d_i \geq 0$. A continuous, strictly concave objective over this compact convex set has a unique maximizer. Lagrange multipliers give the lower-clipped scaled report stated in the method; λ can be found by monotone bisection (the report remains unique when the multipliers are not). Each payoff increment is bounded below by $-\ell(p, r)$; summing proves $z_i(k) \geq -u_i$. The acquisition and trading checks preserve $u_i \leq \kappa(W_i - c_i)$, proving solvency. Across every trade by every agent, the payoff terms telescope, irrespective of order, to $\sum_i z_i(y) = b \log[p^{\text{final}}(y)/p^0(y)] \leq b \log K$ for uniform p^0. For nonuniform p^0, replace the bound by $b \log(1/\min_k p^0(k))$. □

Risk-neutral myopic expected payoff under belief q is $b \sum_k q(k) \log[r(k)/p(k)] = b[\mathrm{KL}(q \,\|\, p) - \mathrm{KL}(q \,\|\, r)]$. Thus unrestricted truthful reporting is optimal when q is feasible; otherwise the unique constrained solution minimizes $\mathrm{KL}(q \,\|\, r)$ over the feasible set. This characterization covers risk-neutral myopic incentives; log utility and non-myopic strategies require separate analysis.

Lemma 3 (Conditional external acquisition expenditure)**.** *If one common measure $\mathbb{P}$ governs all expectations, all variables are integrable, information payments are external, and every agent satisfies $\mathbb{E}_{\mathbb{P}}[z_i(Y) - c_i] \geq 0$ over its full strategy, then $\sum_i \mathbb{E}_{\mathbb{P}}[c_i] \leq \mathbb{E}_{\mathbb{P}}[\sum_i z_i(Y)] \leq b \log K$.*

Proof. Sum the participation inequalities under the common measure and apply the pointwise maker-payout bound. The argument uses a common probability measure for all expectations. Internal evidence transfers cancel in a complete ledger; house endowments and normalization subsidies are additional funding sources outside this bound. □

B.4 DISAGREEMENT AND LIMITED MANIPULATION STATEMENTS

Lemma 4 (Disagreement decomposition)**.** *For positive report weights summing to one and $\bar{q} = \sum_i \widetilde{w}_i q_i$,*

$$\sum_i \widetilde{w}_i \mathrm{KL}(q_i \,\|\, p) = \underbrace{H(\bar{q}) - \sum_i \widetilde{w}_i H(q_i)}_{J} + \underbrace{\mathrm{KL}(\bar{q} \,\|\, p)}_{R}.$$

Moreover $0 \leq J \leq \min\{H(\widetilde{w}), \log K\}$, and $J = 0$ iff all positively weighted reports coincide.

Proof. Insert $\log(q_i/p) = \log(q_i/\bar{q}) + \log(\bar{q}/p)$ and sum. The first term is the generalized Jensen–Shannon divergence. For $A \sim \widetilde{w}$ and $Y \mid A = i \sim q_i$, it equals $I(A; Y)$, proving the bounds. Strict concavity of entropy gives the zero condition. □

If both agents report $(0.5, 0.5)$ and the final LMSR price is $(0.9, 0.1)$, then $J = 0$ while $D = R = 0.510826$. This example separates report disagreement from the market residual. Neither J nor R varies monotonically with error across all cases.

Lemma 5 (Capital-limited direct influence)**.** *For a fixed candidate set and honest sealed pool with stake M and price p, a coalition adding stake s with pooled report σ produces $\mathrm{TV}(p', p) = \frac{s}{M+s} \mathrm{TV}(\sigma, p) \leq s/(M+s)$. Splitting a fixed stake among identical copies leaves p' exactly unchanged. In LMSR, for a coalition's contiguous block of moves with total conservative collateral charged at most U, $p^{\text{after}}(k) \geq p^{\text{before}}(k) e^{-U/b}$.*

Proof. The sealed identity follows by subtracting p from $p' = (Mp + s\sigma)/(M+s)$. For LMSR, each move satisfies $\log[p^{j-1}(k)/p^j(k)] \leq \ell_j/b$. Summing a contiguous block telescopes, and $\sum_j \ell_j \leq U$ gives the result. For interleaved trades the sum still bounds the coalition's own direct log-price contributions, but no longer equals the total before/after log-price difference. □

These bounds quantify direct influence under fixed candidate sets. The core protocol uses authenticated fixed seats, fixed turn quotas, identity-independent candidate truncation, and no on-demand endowment. If a surrounding system permits identity splitting to obtain extra turns, proposals, evidence routes, or subsidies, those channels require separate controls. Collusion, deceptive multi-round trading, and interference with others' responses remain possible (Ostrovsky, 2012; Bara, 2026).

Two counterexamples to overinterpreting wealth. With finite parimutuel impact, another equal-stake trader reporting $(0.5, 0.5)$ and one's true belief $(0.6, 0.4)$, full proportional allocation $(0.6, 0.4)$ has expected log growth 0.005094, while $(0.55, 0.45)$ gives 0.006285. The lower allocation yields higher strategic log growth in this case. At fractional stake $\kappa = 0.5$, two realized report-to-price ratios 4 and 0.2 give cumulative log ratio $\log 0.8 < 0$ but wealth factor $2.5 \times 0.6 = 1.5 > 1$. Positive wealth growth can therefore coexist with worse cumulative log-loss.

C Design Choices and Baseline Implementations

The main text specifies the shared-account loop. Here we define the controls that make its causal interpretation testable: identical access and accounting, a surgical separation of acquisition from influence, and optional population changes evaluated outside the core system.

The account-level effect of a purchase on current influence and the next task's budget is illustrated in Figure 2.

C.1 Defaults and resource semantics

The core population has 16 immutable persona identities, novice wealth $W_0 = 100$, $\kappa = 0.5$, $b = 50$, at most two LMSR rounds, and no rent or renewal. Four posted service levels cost $(2, 4, 8, 16)$ credits; a two-service bundle costs the sum of its component prices, with no hidden discount. These credits model scarcity; the token meter bills actual service generation and consumption. A cheap tier need not be informative, and a high tier need not be correct. A purchaser may revisit the menu after observing a service; the number of decisions is capped at four. The default reserve smoothing is 10^{-4} per class. New model calls use a total input-plus-output cap; reasoning tokens are a subset of output tokens, not an additional charge.

Identical evidence is cached: provider generation is billed once, and each buyer's reading is billed separately. Cached input tokens are reported separately for monetary pricing and still count toward the raw token meter. Peer rationales are generated in the sealed stage, so their generation is already paid; a new rationale incurs a new generation charge. Code execution depends on the action and is metered as an executed service. Unavailable, timed-out, or malformed services retain actual incurred costs and generate an explicit failure receipt.

C.2 A surgical shared versus separate budget comparison

Both arms begin with the same wealth state W_i, personas, labels, service menu, fee schedule, query ordering, and global compute cap. In the *shared* arm, the influence balance is $B_i = W_i - c_i$, with update Equation (5). In the *separate* arm, the controller sees the same remaining acquisition allowance $W_i - c_i$, while purchases leave prediction weights unchanged. Sealed aggregation uses $p = \sum_i w_i q_i$ and updates reliability by Equation (3). Acquisition and reliability are distinct state variables; reliability units are not additional spendable tokens. Both arms use the same normalization to a total of 1600 credits before the next task, so they differ in the opportunity cost imposed on current and future influence.

For the matched LMSR comparison, both arms cap cumulative evidence purchases at $(1-\kappa)W_i$. The separate arm keeps a fixed prediction-collateral allowance κW_i independent of realized purchases; the shared arm uses $\kappa(W_i - c_i)$ as in the main mechanism. Thus evidence plus reserved loss stays within initial capital in either arm, while only the shared arm charges purchases against predictive influence. Both arms start from the same purchased-information permissions and token cap. Separate prediction accounts update by realized trading payoff, while information payments enter an external acquisition ledger; a prediction-account-only ledger therefore omits acquisition costs. The primary comparison is sealed: shared 64.8 versus separate 64.0 on Qwen 2Wiki. The selected-mode comparison uses the

matching LMSR separate-budget arm on evidence tasks (also 64.0). These are distinct experiments even when their rounded numbers coincide.

Matched-report replay versus full closed-loop comparison. A replay fixes candidate lists, persona versions, purchases, probability reports, labels, and costs, then applies alternative pooling rules. It isolates the aggregator conditional on a realized trajectory. Equal-weight, $\kappa = 1$ Bayesian, and fractional updates use the same trace. With the same cost adjustment, fractional replay is *exactly Scrip sealed*, so the numerical difference is zero. A full closed-loop run lets acquisitions react to each arm's evolving state and tests the combined resource-allocation hypothesis. Together, the replay and closed-loop comparison isolate the operator and evaluate the full system.

C.3 Strong controls

Online pools. Bayesian weighting multiplies each weight by the realized report likelihood and normalizes. Fractional weighting uses Equation (3). Hyperparameters, smoothing, and optional fixed-share mixing are tuned only on validation items. Both receive all resolved training labels and the same acquisition controller. A uniform-pool arm keeps acquisition fixed. Random and round-robin acquisition use the same menu and per-task cap, matching their access and resource allowance.

Calibrated stacking. Candidate identities change across questions. The stack therefore shares coefficients across candidate slots: $g(k) = \beta_0 \mathbf{1}\{k = \perp\} + \sum_i \beta_i \log q_i(k)$, followed by softmax and a held-out temperature. For binary forecasting this is logistic stacking on agent logits with an intercept. Regularization is tuned on disjoint validation items. A separate calibration fit uses no test labels; every method receives the same calibration option.

Communication. Debate and its sparse version, MoA, and Judge can select services through the same routing interface. InfoDelphi retains public/private evidence routing, rationale sharing, and confidence-weighted logit aggregation; fixed-evidence and purchasing-enabled variants are distinguished. LMAD localizes conflicting steps; DEAR includes its learned relationship controller; Consilience includes its calibration and calibration-label costs; Meta-Moderator is a supplementary learned-controller arm (Figure 28). The PolySwarm-style internal control omits its external-market blend on 2Wiki, which has no external market price. We adapt published methods to our task and report these results under the adapted protocol. Galanis-style LMSR uses fixed initial private evidence without shared acquisition costs, while persistent-wealth LMSR adds history-dependent capital without the coupled purchase debit.

Single-agent compute. A strict Single-CoT run uses one agent to produce a single continuous answer without a peer rationale. Its per-call output allowance is sized to use the task cap productively. A separate Single-Iterative run may self-revise and invoke tools, with each pass billed. Report the actual lower-cost Pareto point when an endpoint cannot use the full cap. The all-tier Single variant is reported separately because longer context can reduce accuracy. An agent reading generated peer rationales is an ensemble-derived variant.

C.4 Optional renewal, concentration, and identity

The renewal arm uses rent $\min\{0.2, U_i\}$, retirement below one credit, and one-for-one replacement so the population remains 16. Amendment occurs with probability 0.6, otherwise a fresh persona is generated. Every 25 training tasks, a mutated richest-persona child replaces the poorest surviving persona; duplicate replacement in the same step is suppressed. Every child receives a new identity and a logged 100-credit house endowment; all retired balances return to the house. Any normalization after these transactions is logged separately. Label-conditioned amendment uses only the parent's last five training failures. The fixed-persona arm isolates reliability learning from prompt changes.

Controls include all-tier context, no rescaling, no wealth carryover, static learned weights, fixed-share mixing under distribution shift, and affordable-source caps when provenance overlaps. Error correlation uses paired correctness indicators across common held-out items, stratified by task difficulty where possible. Persona text and JS diversity describe textual variation; outcome covariation measures statistical dependence. Effective sample size $N/[1 + (N - 1)\rho]$ summarizes exchangeable correlation.

D EVALUATION PROTOCOL AND REPRODUCIBILITY

This appendix supplies the complete numerical record behind the question-led figures in the main text, together with dataset, run, and statistical protocols. Exact finite-state calculations and the controlled specialist simulation are identified separately from LLM benchmark runs.

D.1 COMPLETE PERFORMANCE AND ATTRIBUTION RESULTS

Figure 1 selects representative controlled-task comparisons for readability. The result atlas below retains all 19 methods, both market modes, and the validation-locked composition while replacing dense matrices with task-level rankings, direct labels, and uncertainty plots. The original `tabs/` records remain the numerical audit source for every plotted point.

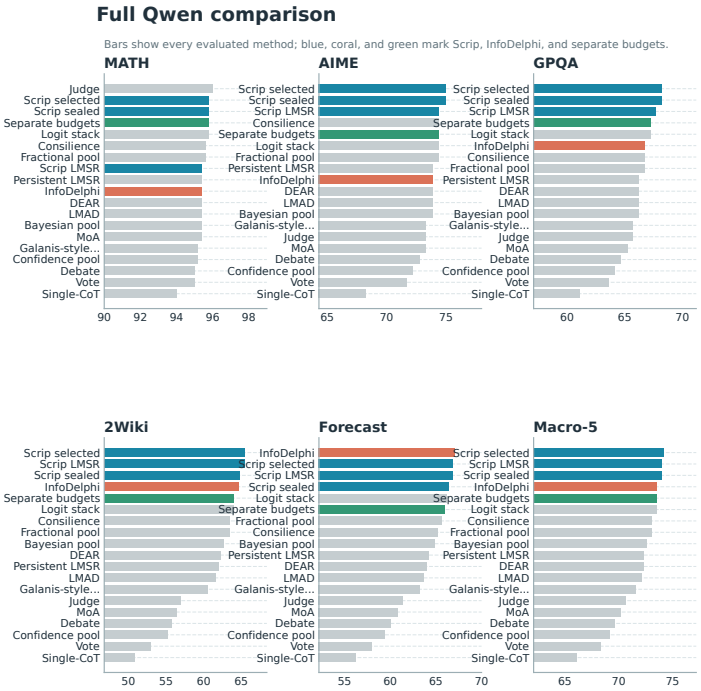

Figure 5: **Full Qwen comparison.** Every method and benchmark is retained; the highlighted bars identify Scrip, InfoDelphi, and separate budgets.

The Qwen panel preserves task-level variation across the full method set; the next panel checks how those rankings carry over to GPT.

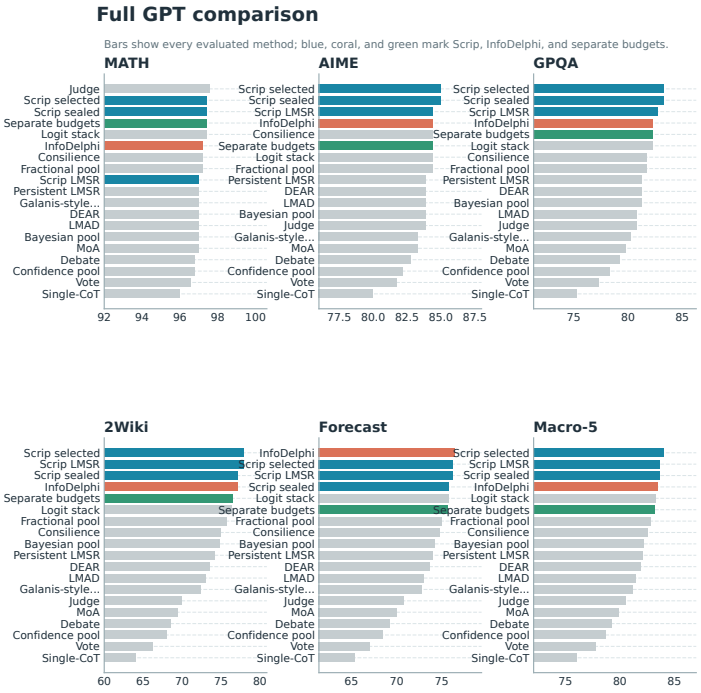

Figure 6: **Full GPT comparison.** The same complete matrix is shown as six task-level ranked panels, with direct labels on the strongest matched-access controls.

The paired view below focuses on shared-versus-control differences and their uncertainty across items and training streams.

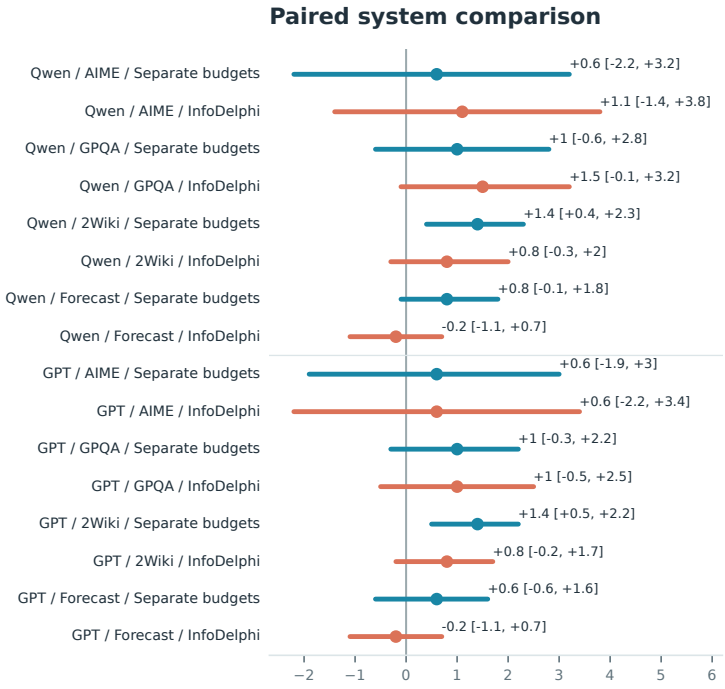

Figure 7: **Paired system differences.** Each point is Scrip minus the comparator in percentage points; horizontal bars are the supplied 95% crossed stream/item intervals and the vertical rule marks zero.

These intervals locate the comparisons with the clearest paired separation; the ablation then separates acquisition, weighting, and budget coupling.

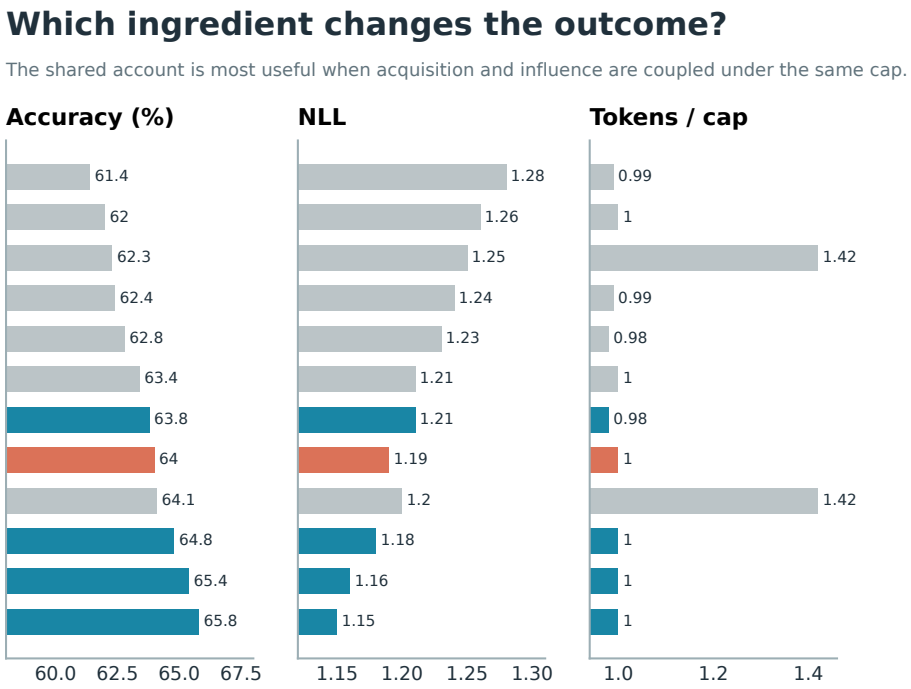

Figure 8: **Economic ingredients.** Accuracy, NLL, and cap usage are shown in aligned panels so the performance gain is read together with probability quality and resource use.

The matched replay holds the realized trajectory fixed to identify the sealed aggregation rule.

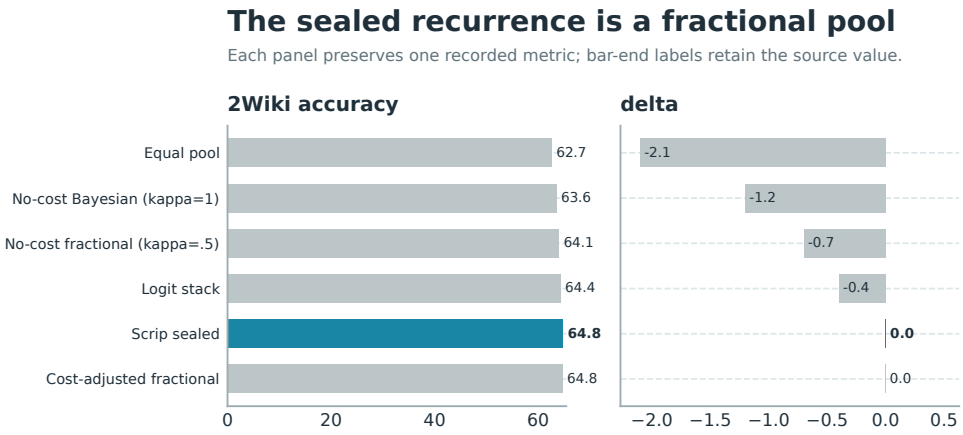

Figure 9: **Matched-report replay.** Holding candidates, purchases, reports, labels, and fees fixed makes the sealed recurrence coincide exactly with the cost-adjusted fractional pool.

The original shared-account interval provides the benchmark-scale contrast before evidence affordability is varied.

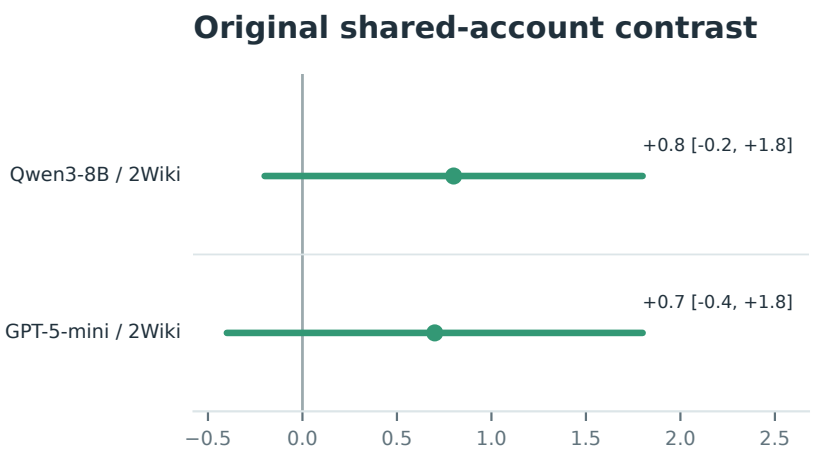

Figure 10: **Original shared-account contrast.** The intervals show the sealed shared-minus-separate accuracy difference on the original 2Wiki evaluation.

D.2 Expanded evidence-allocation measurements

The motivation diagnostics in Figure 11 reweight or route the same reports with retrospective labels. Gold-support holder weighting sets nonholders' weights to zero before renormalization; the correct-agent router chooses a correct report if one exists. These diagnostics quantify the opportunity available from acquisition under retrospective labels.

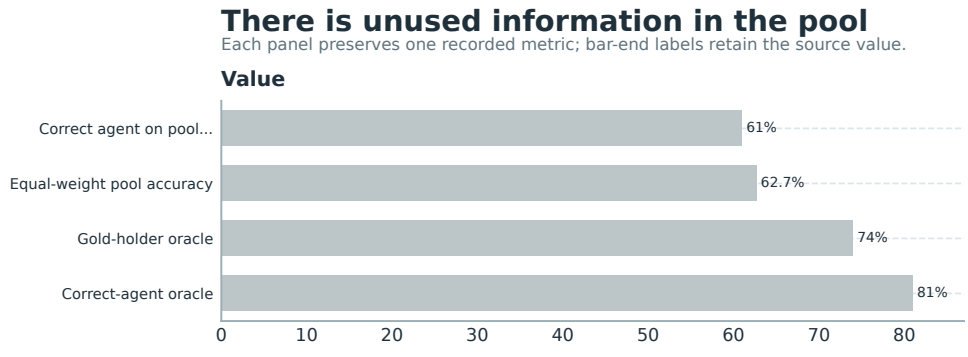

Figure 11: **Unused information in the pool.** The bars compare the observed equal-weight pool with retrospective oracle reallocations; the oracle rows serve as diagnostic ceilings for interpreting acquisition.

The additional Qwen controlled 2Wiki study varies the evidence affordability k across five levels while keeping sealed aggregation fixed in both arms. Its paired accuracy summaries appear in Figure 12; the mean correctness correlations and the price-heterogeneity spending shares appear in Figures 13 and 14. The expanded baseline comparison reports $b = 118$ stream-item observations on which only the shared arm is correct and $c = 78$ on which only the separate arm is correct across 5,000 paired observations. Their difference is $(b - c)/5000 = 0.008$, or $+0.80$ percentage points, with a reported 95% interval $[+0.25, +1.35]$. A two-sided exact binomial test of the discordant pairs gives $p = 0.0052$; because the five streams reuse question identities, the paired interval with item/stream resampling is the inferential summary. The affordability scan separately reports $[+0.2, +1.4]$ at $k \geq 6$ for the same point difference; the two intervals are kept with their respective analyses rather than pooled. The five-seed summaries report variability in scarcity, price heterogeneity, and retrieval.

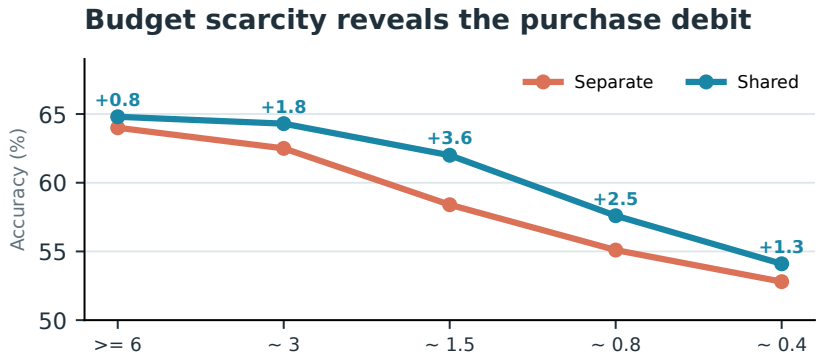

Figure 12: **Scarcity exposes the purchase debit.** Shared and separate budgets are plotted across all five affordability regimes, with the paired gain written at each point.

The gain peaks at intermediate affordability; the correlation panel describes how evidence access accompanies changes in error dependence.

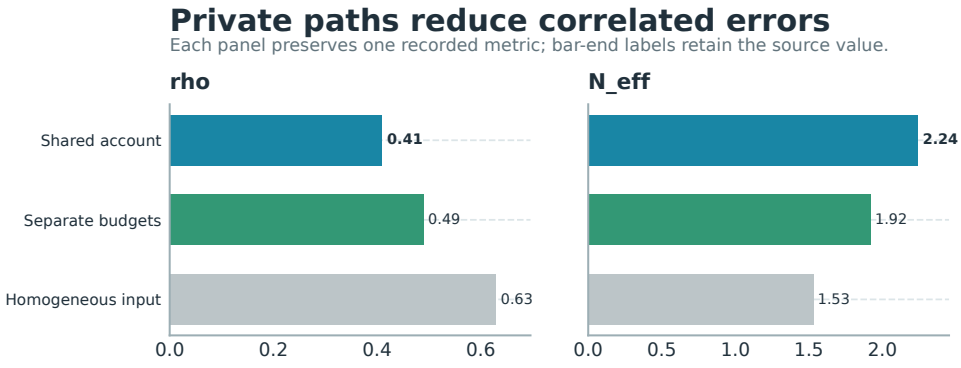

Figure 13: **Different evidence paths.** Mean pairwise correctness correlation and its exchangeable effective population are shown side by side for homogeneous, separate, and shared access.

The next comparison links heterogeneous prices to the composition of purchases and the resulting accuracy.

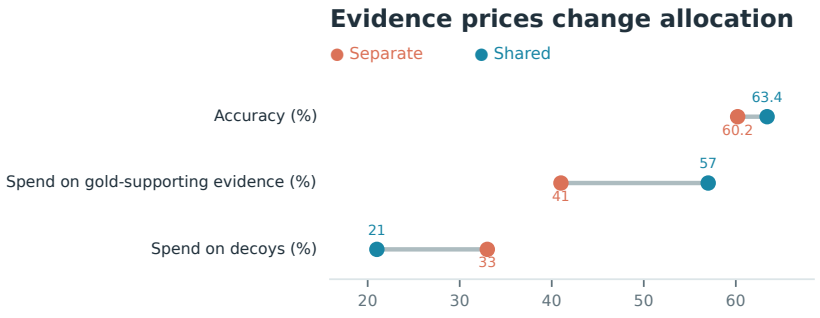

Figure 14: **Price heterogeneity changes allocation.** The connected endpoints show how shared accounting increases spending on gold-supporting passages and reduces decoy spending while improving accuracy.

The adaptive-seller stress test measures how the same buyer-side rule performs when prices respond strategically.

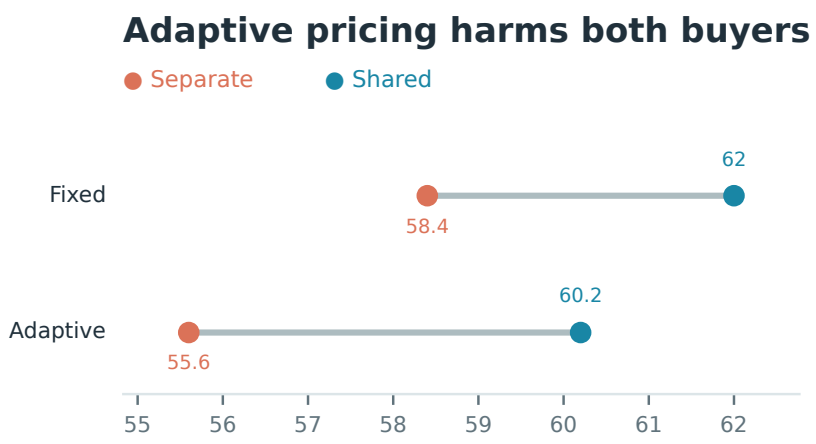

Figure 15: **Adaptive seller pricing.** The same scarcity regime under fixed and adaptive pricing shows that seller incentives remain active under buyer-side accounting.

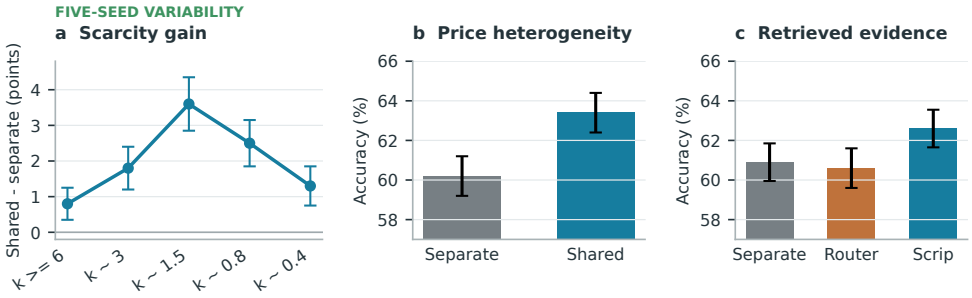

Figure 16: **Five-seed variability.** Points show the reported means; error bars show the across-seed standard deviations.

The seed summaries complement the paired item/stream intervals above by showing variability across independently trained account trajectories.

Table 1: **Five-seed results.** Means and across-seed standard deviations for controlled 2Wiki and the expanded evidence studies. Variance is the square of the displayed standard deviation, in squared percentage points.

Study	Condition	Mean	Seed s.d.	Seed var.
Overview	Judge accuracy	57.0	0.55	0.3025
Overview	Separate accuracy	64.0	0.70	0.4900
Overview	Scrip sealed accuracy	64.8	0.65	0.4225
Overview	Scrip LMSR accuracy	65.4	0.60	0.3600
Scarcity	$k \geq 6$, shared − separate	+0.8	0.45	0.2025
Scarcity	$k \approx 3$, shared − separate	+1.8	0.60	0.3600
Scarcity	$k \approx 1.5$, shared − separate	+3.6	0.75	0.5625
Scarcity	$k \approx 0.8$, shared − separate	+2.5	0.65	0.4225
Scarcity	$k \approx 0.4$, shared − separate	+1.3	0.55	0.3025
Price	Separate accuracy	60.2	1.00	1.0000
Price	Shared accuracy	63.4	1.00	1.0000
Retrieval	Separate accuracy	60.9	0.95	0.9025
Retrieval	Router accuracy	60.6	1.00	1.0000
Retrieval	Scrip accuracy	62.6	0.95	0.9025

The original fixed-price 500-question retrieval diagnostic appears in Figure 29. The expanded 1,000-question non-oracle result in Figure 30 is a separate evaluation and includes a spend-matched relevance router. The route-only comparison at $k \approx 1.5$ (60.8% versus 62.0%) comes from the controlled scarcity protocol; the non-oracle retrieval chart reports 60.6% versus 62.6%. Adaptive seller pricing reduces separate-budget accuracy by 2.8 points and shared-budget accuracy by 1.8 points relative to fixed prices. The shared advantage is 4.6 points in this stress test, while lower absolute accuracy in both arms shows that buyer-side accounting leaves seller incentives active.

D.3 Complete probability and information-structure results

The forecasting table uses scalar binary Brier, while the family-balanced table uses vector Brier. The exact information table integrates finite-state distributions and therefore requires no sampling interval. These complementary metrics explain why an unchanged decision can coexist with a better or worse probability forecast.

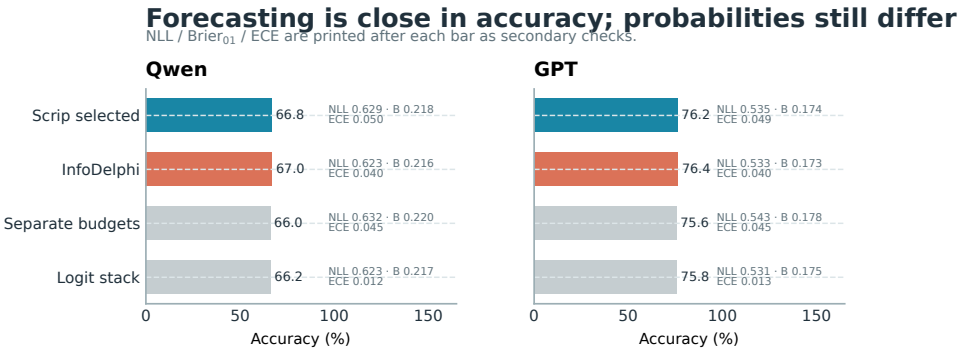

Figure 17: **Forecasting as probability estimation.** Accuracy is the bar length; NLL, scalar Brier, and ECE are printed alongside each method so close decisions can be compared with their probability quality.

The calibration panel separates reliability from accuracy and proper scoring performance.

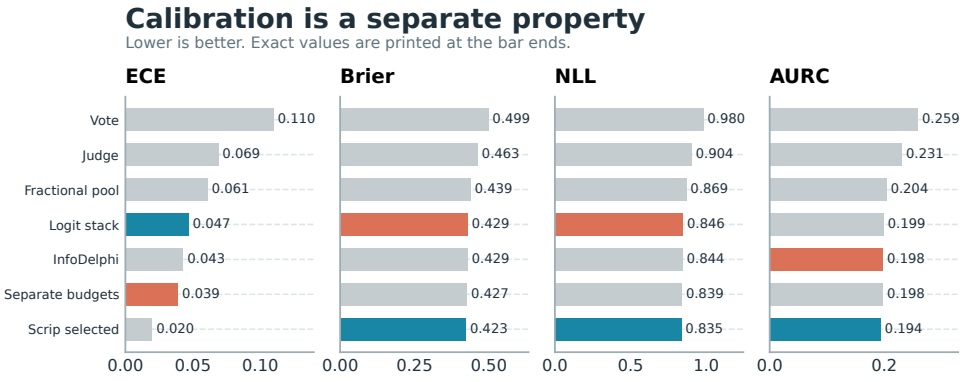

Figure 18: **Calibration is a separate property.** Four lower-is-better probability metrics are shown as aligned method rankings for direct comparison.

The exact finite-state tests isolate how dependence, misleading reliability, and complementary information affect aggregation.

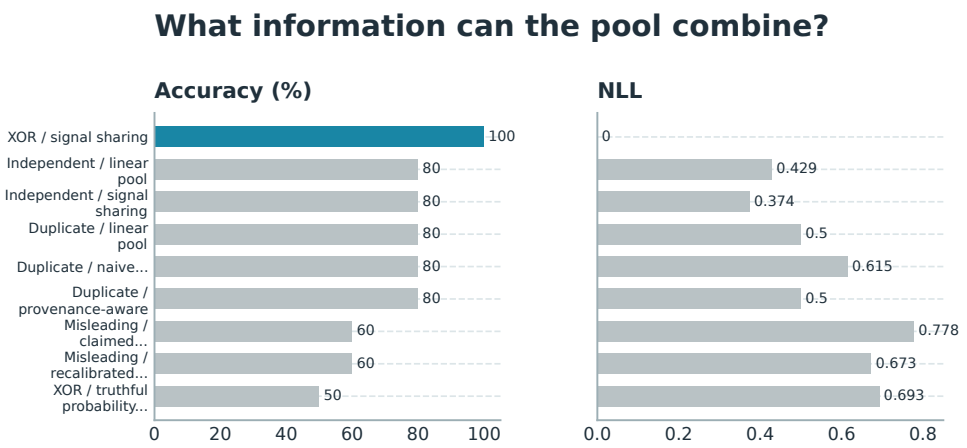

Figure 19: **Information structure determines aggregation.** Accuracy and NLL are separated by information condition, making duplicate evidence, misleading reliability, and XOR complementarity visible without a numeric matrix.

The next panel quantifies paired uncertainty in probability quality.

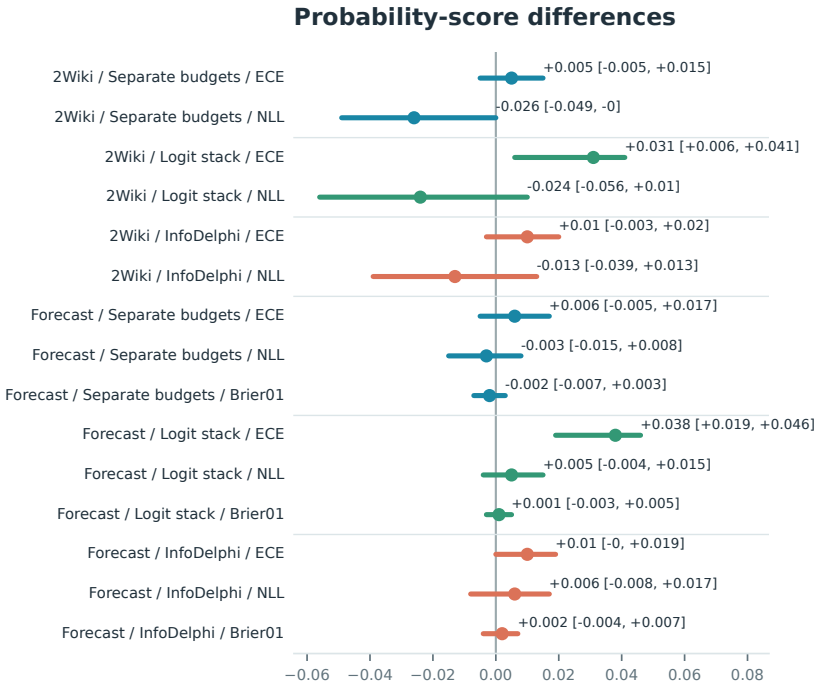

Figure 20: **Paired probability-score differences.** Scrip minus comparator for ECE, NLL, and binary Brier; the zero rule separates evidence for an improvement from overlapping uncertainty.

The absolute panel returns to accuracy and shows its uncertainty across tasks and backbones.

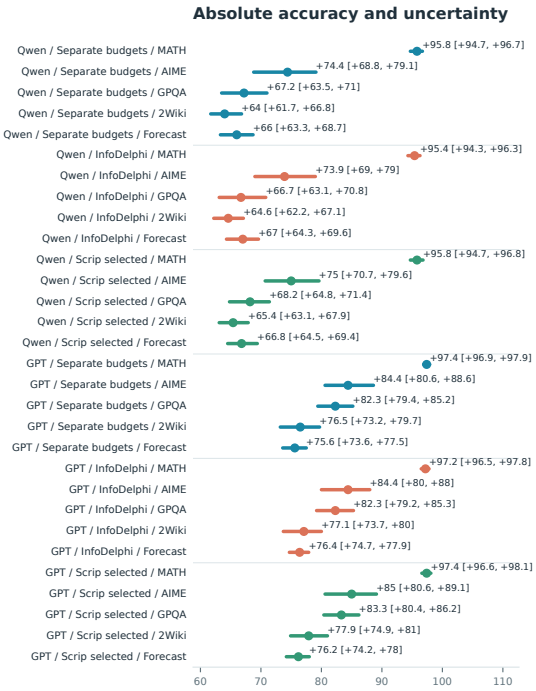

Figure 21: **Absolute performance uncertainty.** Accuracy points and 95% intervals for the central systems across tasks and backbones.

D.4 Dataset separation and provenance

Mathematics. MATH-500 is the held-out test set; AIME has 60 unique items across 2024 and 2025. Training and validation draw from separate MATH training IDs, excluding exact and normalized-text matches to either test set. Symbolic answer equivalence is applied before scoring. The 60 unique AIME items are the units of inference; repeated decodes estimate within-item variation. Candidate construction receives no answer key.

GPQA. GPQA-Diamond (198 items) is a subset of GPQA main (448), leaving 250 non-Diamond items for training and validation. We order their canonical IDs by a public salted SHA-256 hash, assign the first 200 to training and the remaining 50 to validation, and sample each 300-step training stream *with replacement* from the 200 training items. The 198 exact IDs and normalized-question hashes are excluded before sampling; alternate dataset revisions are matched by content as well as IDs. Amendment and calibration use only training and validation labels.

2Wiki. The original data pool consists of 500 disjoint training questions, 100 validation questions, and 500 evaluation questions from designated source splits. The expanded Qwen evidence studies use 1,000 evaluation questions. A training stream has 300 draws from the training pool; pool size and stream length are different quantities. The controlled variant removes gold supporting paragraphs from free context and splits them across paid services with distractors. Descriptions expose length, language, and source type but no gold relevance, supporting-fact flag, or answer string. Shard assignment is fixed by item/agent seeds. Every agent can eventually buy the same services, subject to payment; private free shards differ initially. The original non-oracle variant uses a frozen BM25-ranked paragraph pool and a query-based reranker that never receives supporting-fact labels. Gold support is used only for evaluation. Exact-match answer accuracy is primary; candidate coverage is reported independently.

Forecasting. The primary track uses binary events only; categorical events require a separate class-count-stratified analysis. Event groups, including multiple questions about one underlying event, remain in one split. Events are ordered by resolution time within documented temporal blocks, preserving training/validation/test boundaries. Records include source question ID and URL, question creation time, prediction as-of time, resolution time, outcome, event-cluster ID, snapshot hashes,

document publication and retrieval times, and model checkpoint/API snapshot identifier. Content dated after the as-of time is unavailable to the system. For the post-cutoff track, event resolution follows the documented model-training cutoff and the as-of time precedes resolution. The evaluated binary set contains 250 positive and 250 negative events.

D.5 Run design and measurement

Each original method uses five independently initialized training streams, each with 300 resolved draws and the same labels within a paired stream. Three decoding seeds are averaged within each original test item. The expanded Qwen study reports 1,000 items across five streams. Validation tunes $\kappa \in \{.25, .5, 1\}$, $b \in \{10, 50, 200\}$, rounds in $\{1, 2, 4\}$, acquisition policy, temperature, and abstention ranking within the same tuning cap. The selected configurations are frozen before testing; the full mode grid is still reported. Test items use independent snapshots of the selected training state. A population adaptation, router-training, or calibration operation consumes training/validation resources even without a gradient update.

Backbone identifiers in the result display denote model families. The run record distinguishes weight revisions, tokenizer versions, reasoning settings, API snapshots, serving versions, and random seeds. Elicitation uses temperature 0.7 and deterministic canonicalization when available. Reasoning backbones receive per-task output caps and logged stop reasons. Invalid JSON triggers one charged repair call; a second failure yields a uniform report with a failure flag. If no valid proposal remains, the system returns abstention. All failures stay in the denominator. Tool errors and timeouts are billed by actual usage. Truncated calls account for 1.7%, first-pass JSON failures for 0.6%, and retries for 1.1% in the operational profile.

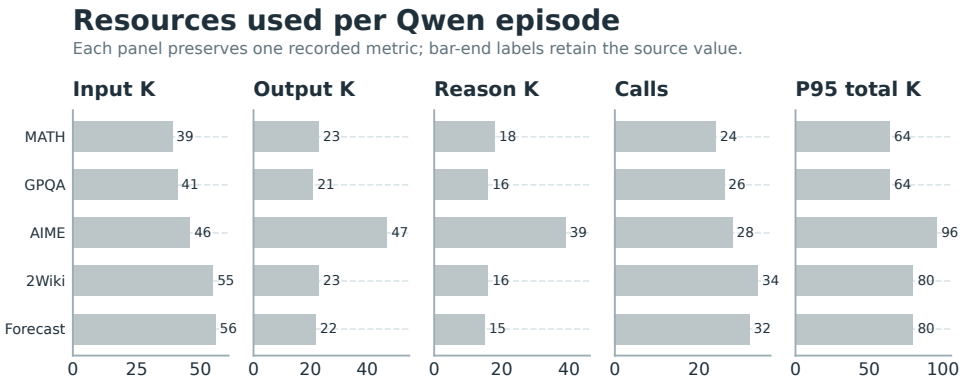

Figure 22: **Episode resource meter.** Input, output, reasoning, calls, and the P95 total are kept together as per-task operating profiles.

Token profiles describe per-task usage; the next panel adds latency, cost ratio, and retry rate.

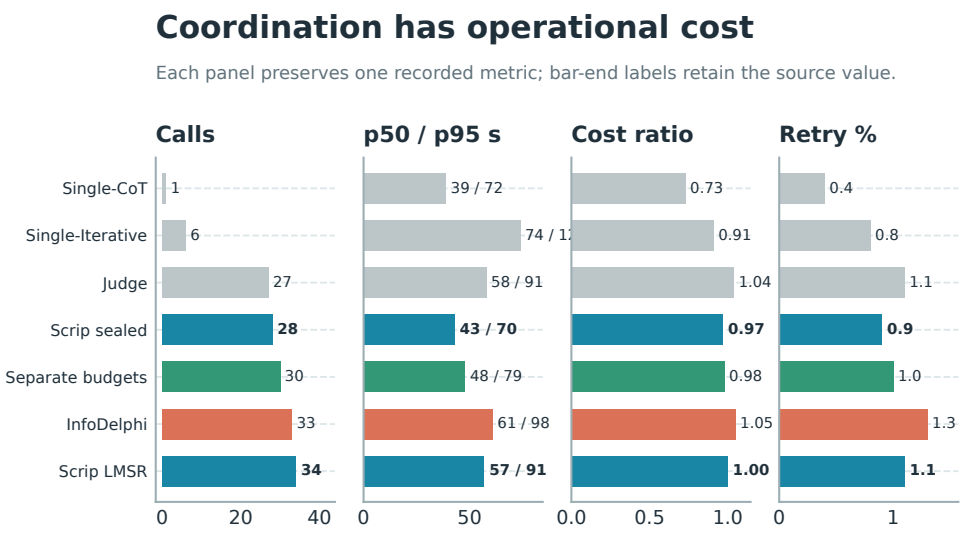

Figure 23: **Operational cost of coordination.** The chart separates median and P95 latency while retaining cost ratio and retry rate in the direct labels.

Training cost and the audited wealth ledger complete the resource account, including adaptation overhead.

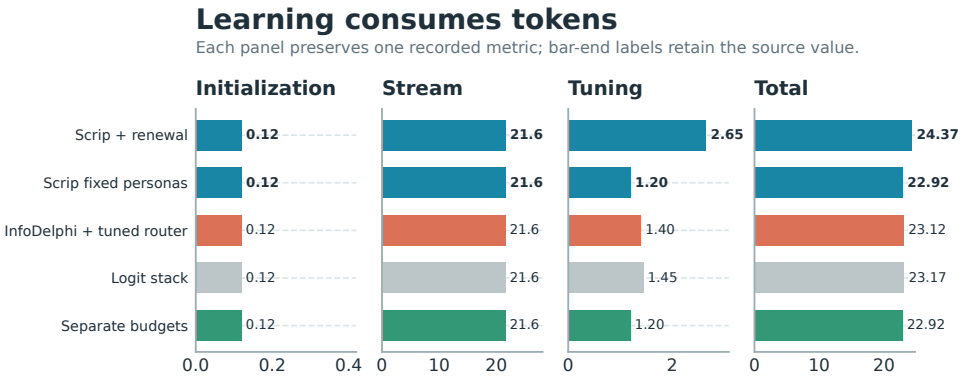

Figure 24: **Cost of adaptation.** Initialization, training stream, controller tuning, and total are shown as aligned metric panels so amortization can be read without a dense accounting table.

The ledger traces how opening balances, transfers, payouts, and rent produce closing wealth.

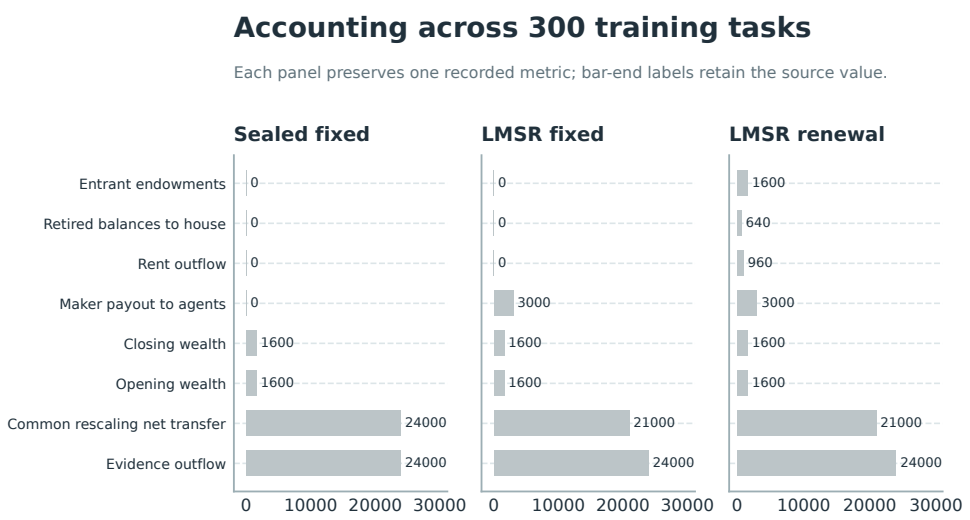

Figure 25: **Financial conservation.** Opening wealth, transfers, payouts, rent, and closing wealth are displayed as aligned measures in the audited training ledger.

Raw tokens count every input and output, with reasoning counted once inside output. Uncached input, cached input, visible output, and reasoning subsets are separated for price conversion. Tool latency and sequential stages are recorded at fixed concurrency 16. Fixed caps define upper resource limits; methods need not consume the full allowance.

For training cost C_{train} and M deployment items, amortized cost is $C_{\text{train}}/M + C_{\text{test}}$. We compare $M \in \{500, 5000, 50000\}$ and include controller fitting and persona generation. With 22.92 million training tokens, these overheads are 45,840, 4,584, and 458.4 tokens per item per training run. Five independent streams provide experimental replication; one trained system is deployed. Renewal adds 1.45 million training tokens.

D.6 Statistics, metrics, and reporting rules

For dataset d, report Acc_d averaged over streams, items, and decoding seeds. Macro-5 is the unweighted mean of the five datasets. Macro-4 is $\frac{1}{4}[\frac{1}{2}(\mathrm{Acc}_{\mathrm{MATH}} + \mathrm{Acc}_{\mathrm{AIME}}) + \mathrm{Acc}_{\mathrm{GPQA}} + \mathrm{Acc}_{\mathrm{2Wiki}} + \mathrm{Acc}_{\mathrm{Forecast}}]$. NLL, Brier, and ECE use the same family weighting when so labeled. The risk plot uses forecasting only and therefore has endpoint 0.332, not the error implied by either macro average.

NLL is $-\log p(y)$; vector Brier is $\sum_k [p(k) - \mathbf{1}\{y = k\}]^2$. Binary event Brier is $(p(1) - y)^2$, half the vector score. Top-label ECE is $\sum_b (n_b/n)|\overline{\max p}_b - \overline{\mathbf{1}\{\arg\max p = y\}}_b|$ on ten fixed-width bins. Empty bins are omitted and their zero support is preserved. Adaptive ECE uses ten equal-count bins, and classwise ECE averages the event-probability ECE over candidate slots. OTHER has its own contract-event interpretation; open-answer contract calibration and task-answer accuracy are not conflated. Coverage ranking uses stable deterministic tie-breaking and all methods use the same coverage grid for AURC. An abstaining policy retains its coverage denominator; at coverage one, all rankings of the same predictions return the same risk.

For paired intervals, decoding seeds are first averaged within each item. Training streams and question IDs are resampled independently as a crossed bootstrap, applying the same draws to every compared method. Forecasting resamples event clusters rather than individual correlated contracts. ECE and proper-score differences are recomputed inside each resample. AIME has 60 item units and GPQA 198; extra decoding seeds do not enlarge these item counts. The primary comparisons are shared versus separate budgets on 2Wiki NLL and forecasting Brier; additional accuracy comparisons are secondary, with multiplicity handled by a predeclared Holm correction. Seed spread is reported separately from item/stream intervals.

Cost frontiers use nondominated observed points; adjacent-point interpolation corresponds to randomized deployment mixtures. An equal-accuracy ratio is defined only on a target achievable by both methods. Each paired resample rebuilds the frontier and solves the target-cost problem; unsupported resamples are counted without extrapolation. Clock time, billed cost, and raw tokens receive separate frontiers.

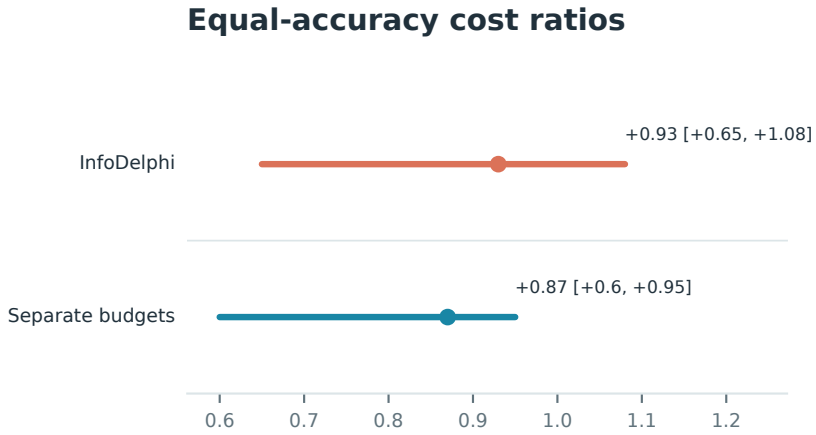

Figure 26: **Equal-accuracy cost.** Ratios below one indicate a lower cap-equivalent cost for Scrip; unsupported resamples are retained in the annotations.

D.7 Additional comparisons and failures

The remaining controls examine the interfaces around the shared account: acquisition quality, retrieval, candidate construction, and resistance to changes in information quality. Together they locate failure modes across the acquisition loop and contextualize system-level effects of the budget rule. The

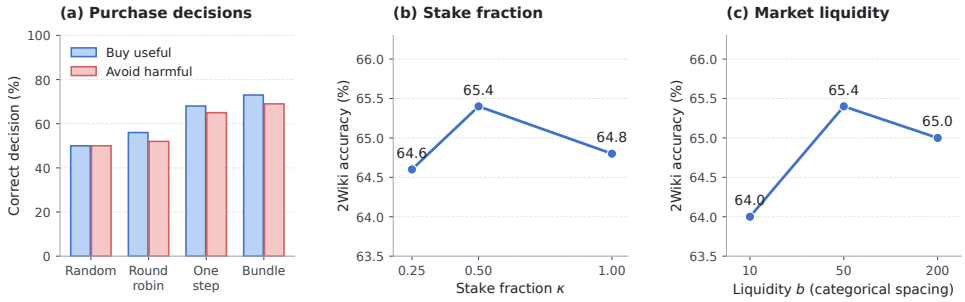

Figure 27: **Purchase-policy diagnostics.** Qwen controlled 2Wiki purchase diagnostics and sensitivity. Bundle selection is compared with simpler acquisition policies; separate sweeps vary stake fraction and liquidity. Connected points indicate tested settings, without confidence bands or an interpolation claim.

forced-buy/no-buy diagnostic starts from a frozen agent state and uses paired decoding seeds in both branches. Both branches share candidates, prior reports, initial wealth, and anticipated external price. We compare held-out log loss and the fractional log-wealth objective after information costs. The complete-market rerun measures downstream effects separately. Since price directly affects collateral, weights, and later access, price perturbations trace downstream sensitivity under endogenous feedback. Diagnostic branch calls are reported as analysis compute.

The acquisition comparison includes prompted one-step, bundle lookahead, random, and round-robin policies; a dynamic-programming policy applies to the tractable toy state space. Confidence-action inconsistency compares verbalized value with realized purchase choices. Conditional error correlation

is reported for uniform versus private evidence and heterogeneous backbones. Source-replication and identity-replication attacks test different mechanisms: fixed identical-report stake splitting is algebraically invariant, while new source paths or turn quotas can alter behavior.

Financial conservation is checked before normalization and renewal. In the ledger, rent and retired balances are house receipts, whereas entrant endowments and recapitalization are house expenses. The total net injection is therefore distinct from maker payout.

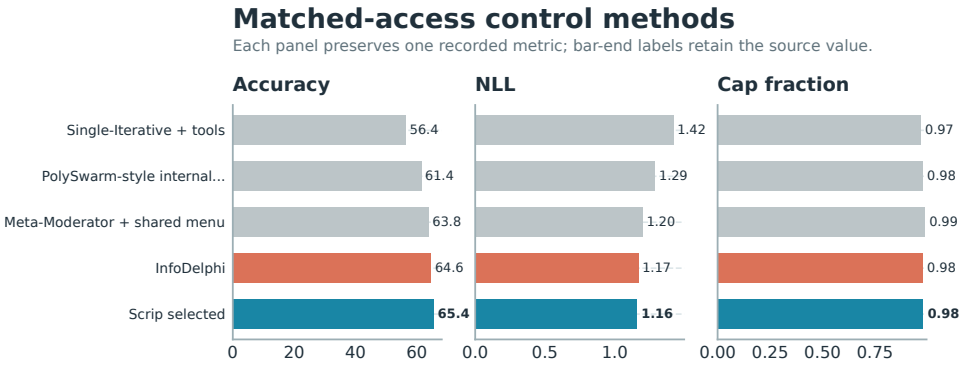

Figure 28: **Matched-access controls.** Single-agent iteration, learned moderation, an internal confidence pool, InfoDelphi, and Scrip share the same compact accuracy ranking.

The original retrieval panel carries the matched methods into a non-oracle setting; the expanded evaluation uses a distinct protocol.

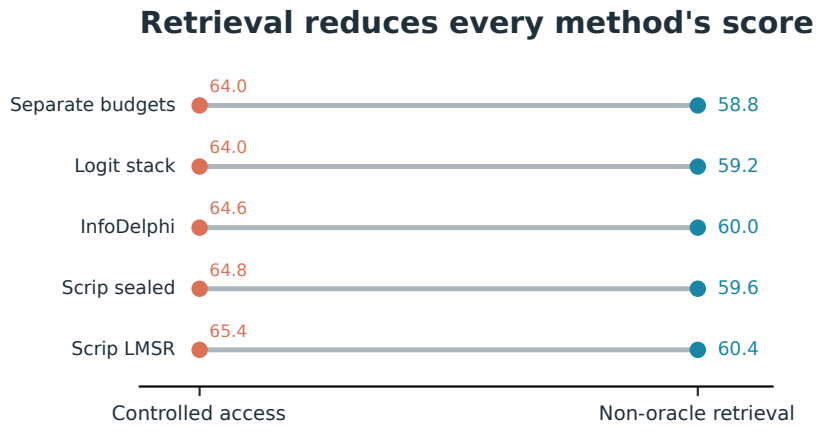

Figure 29: **From controlled access to retrieval.** Each method is connected across the controlled and non-oracle settings; endpoint labels show the accuracy drop.

The expanded retrieval study tests whether the shared-account difference persists over a larger question set.

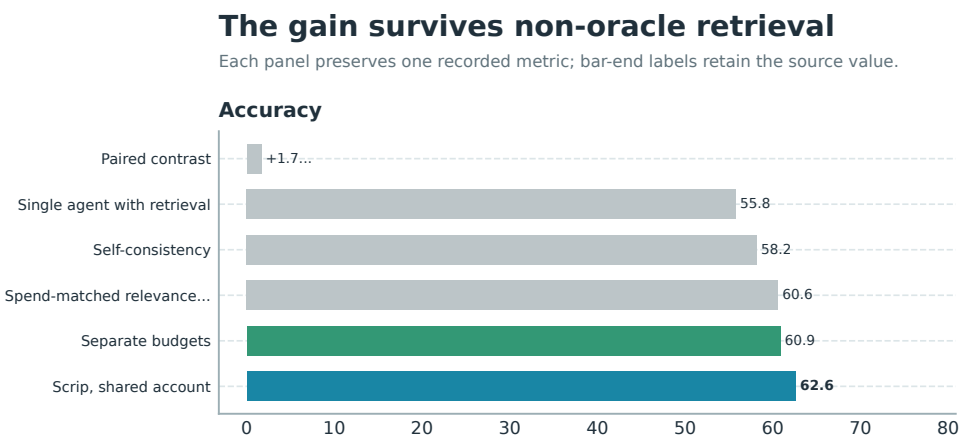

Figure 30: **Expanded retrieval evaluation.** The 1,000-question non-oracle result keeps the shared-account advantage visible alongside the strongest router and separate-budget controls.

Candidate reopening measures answer-space coverage; the recovery count tracks when that expanded set contains the correct answer.

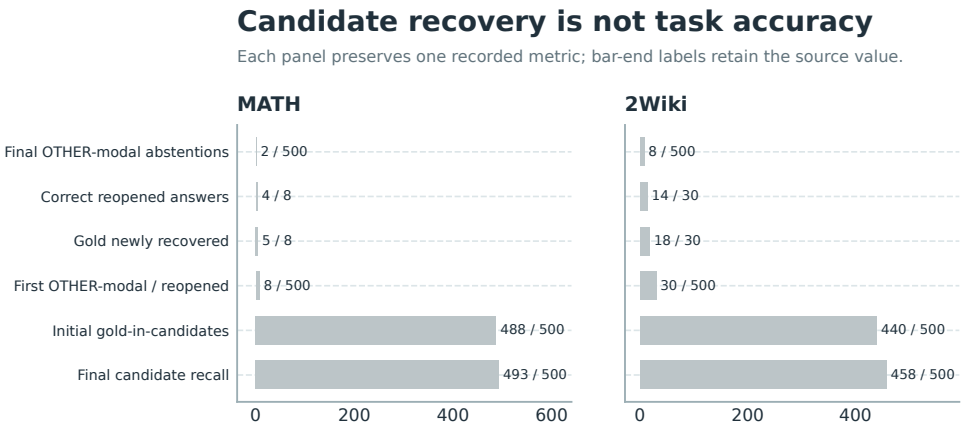

Figure 31: **Candidate recovery is a separate quantity from task accuracy.** Initial coverage, reopened coverage, and the count of correct recoveries are displayed separately.

The final sensitivity panel varies controller and market settings to show which patterns persist across tested choices.

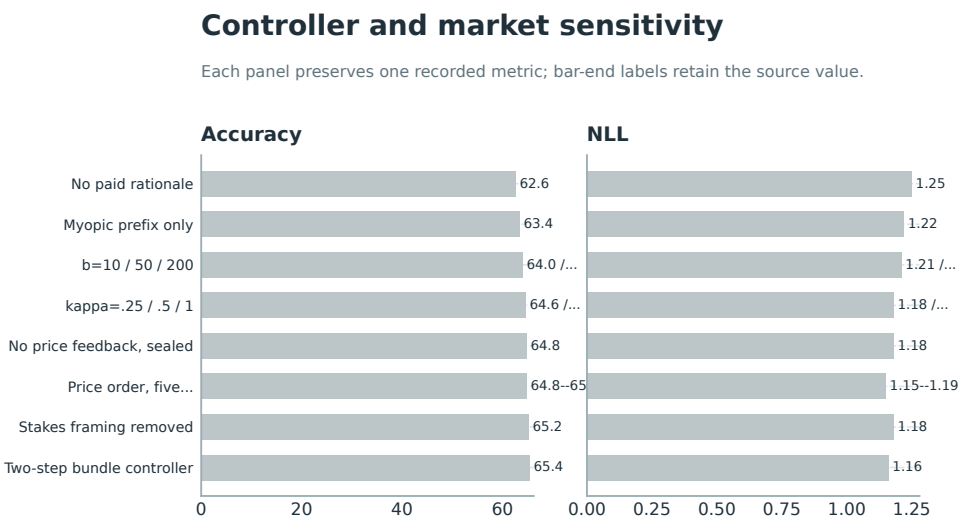

Figure 32: **Controller and market sensitivity.** Each tested choice is shown with its accuracy and NLL, while the connected series preserve the direction of the sweep.

D.8 Probability, resource, and population diagnostics

Accuracy alone leaves probability quality unresolved. This matters for forecasting, where the primary outputs are probabilities, and for selective prediction, where a system may defer uncertain items. Figure 33 links proper scores, reliability, and abstention to assess normalized market prices as probabilistic forecasts.

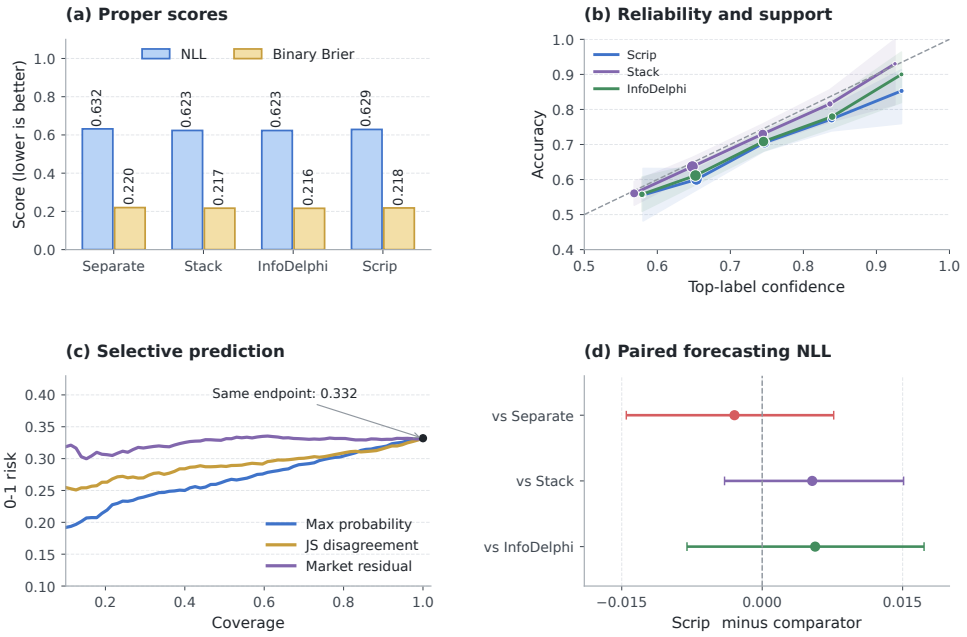

Figure 33: **Probability quality requires several checks.** Qwen binary forecasting: **(a)** NLL and scalar event Brier compare four strong methods; lower is better. **(b)** Top-label reliability uses the same predictions as ECE, with marker area proportional to bin support and crossed-bootstrap bands. **(c)** Three rankings retain different subsets of the same Scrip predictions, so all meet at full-coverage risk 0.332. **(d)** Paired NLL differences and intervals show uncertainty relative to the strongest alternatives.

Calibrated stacking and InfoDelphi remain competitive with Scrip on forecasting NLL and Brier, and the paired intervals overlap zero. Calibration can improve probabilities while preserving the chosen answer. All methods receive the same option of a validation-fitted temperature. Binary event Brier is reported separately from the multiclass vector score, which differs by a factor of two; the linked panels carry the numerical comparison.

The reliability panel retains supported top-label confidence bins, which begin at 0.5 for a binary problem. The selective-risk panel compares maximum probability, report disagreement J, and market residual R on identical outputs. Their common endpoint is an accounting check, while AURC and paired score differences quantify the remaining distinction (Figure 20). Low disagreement can reflect shared error, so we interpret it alongside outcome-based scores.

D.9 Resource use and adaptation

Resource comparisons pair task performance with inference and training costs, since adaptation changes the cost per deployment item. The relevant resource curve is task-dependent. Judge remains



Figure 34: **A resource advantage must survive strong controls and training cost. (a,b)** Measured operating points at three predeclared cap settings; lines denote randomized mixtures between adjacent configurations. Judge retains the MATH advantage. **(c)** At the common 2Wiki target of 64% accuracy, interpolated cap-equivalent costs differ modestly. **(d)** Training overhead per deployment item falls with deployment volume.

stronger on high-budget mathematics, while Scrip's 2Wiki operating points have a modest advantage over the matched acquisition controls (Figure 34). At the common 64% target, adjacent-point interpolation gives about $0.87\times$ the nominal cap for Scrip, $0.94\times$ for InfoDelphi, and $1.00\times$ for separate budgets. These cap-equivalent comparisons summarize randomized mixtures between measured configurations; the cost frontier is encoded directly in the figure and its ratio interval against InfoDelphi includes one.

Learning the accounts also costs computation. The 22.92-million-token fixed-persona training stream adds 45,840 tokens per deployment item at 500 items, but 458.4 at 50,000. The same amortization applies to the learned controls; the operating-cost panels distinguish input, output, reasoning subsets, latency, retries, and optional renewal overhead without repeating them in separate tables.

D.10 POPULATION FEEDBACK AND INFORMATION STRUCTURE

To isolate historical weighting, we execute a controlled specialist simulation where competence is known and deliberately changes. Sixteen fixed experts predict 300 binary outcomes; the accuracies of two eight-expert groups switch from $(0.8, 0.55)$ to $(0.45, 0.8)$ after task 150. All methods receive the same realized reports, with purchases and persona replacement held fixed.

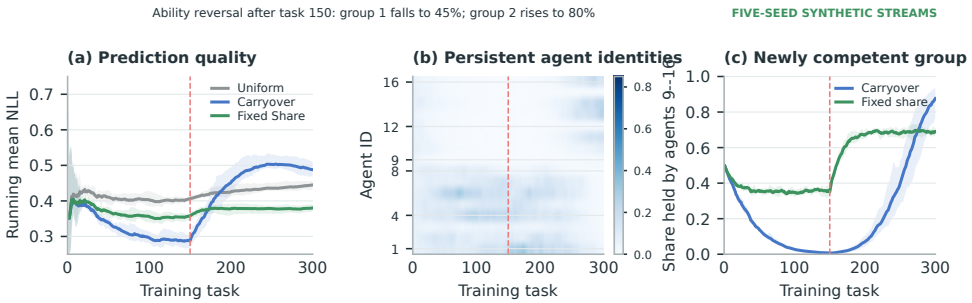

Figure 35: **Past success can become an obstacle after expertise changes.** Five-seed synthetic streams. **(a)** Running mean NLL with the seed envelope for uniform, fractional carryover, and 5% fixed-share weighting. **(b)** A representative wealth-share heatmap follows all 16 persistent identities. **(c)** Group-level shares show recovery of newly competent specialists; the seed envelope keeps the original comparison visible.

Figure 35 shows concentration and recovery as distinct phenomena. Ordinary carryover emphasizes initially successful specialists, then reacts slowly when their skills deteriorate. Across five seeds, fixed-share weighting recovers the newly competent group faster and retains the best second-half accuracy ($92.3 \pm 3.9\%$), compared with $85.9 \pm 4.5\%$ for uniform pooling and $60.5 \pm 2.6\%$ for carryover. Fixed-share mixing preserves a route back for neglected experts.

Exact finite-state cases separate the limits of probability pooling from the benefit of exchanging a missing signal. Exact enumeration exposes why the acquisition policy must distinguish new information from repeated confidence. With two independent 80%-reliable signals, ideal signal sharing improves NLL from 0.429 to 0.374 while accuracy remains 80%. Duplicating a single source adds no information; treating the copies as independent worsens NLL from 0.500 to 0.615. When a source claims 80% reliability but achieves only 60%, correcting that assumption improves NLL from 0.778 to 0.673 (Figure 36).

XOR provides the complementary case. Exhaustive enumeration of its four states gives 50% accuracy and NLL $\log 2$ for truthful answer-probability exchange, versus 100% and zero NLL for an exact solver after signal sharing. A priced rationale can carry the missing signal, so useful communication is part of the proposed acquisition mechanism.

The benchmark stress tests examine these mechanisms. Identical evidence erodes the advantage; misleading passages favor InfoDelphi; frozen historical influence underperforms separate budgets after an ability shift. Provenance-aware source caps and fixed-share updates test defenses against these failures. In the original non-oracle condition, Scrip LMSR falls from 65.4% to 60.4%, while the expanded retrieval result is reported in Section 4.4. Reopening raises 2Wiki candidate recall from 440 to 458 of 500 items, but only 14 reopened answers are correct. Finding the right answer space and choosing the right answer remain different problems.

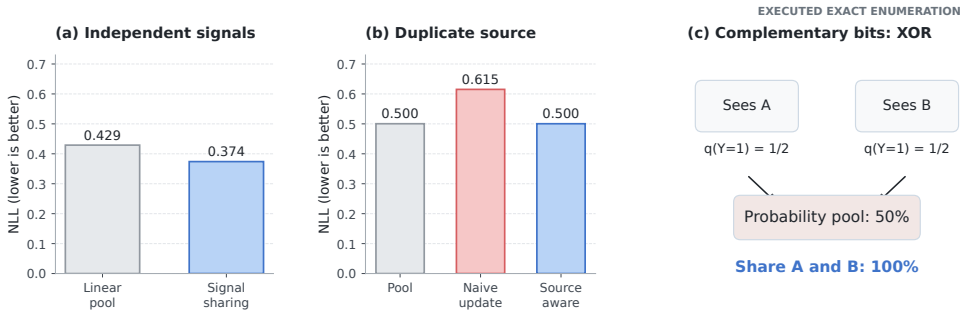

Figure 36: **Information structure determines what aggregation can achieve.** Executed exact finite-state calculations. **(a)** Sharing two independent 80%-reliable signals improves NLL over linear pooling without improving accuracy under the stated tie rule. **(b)** Treating duplicate evidence as independent increases overconfidence; provenance-aware treatment removes that error. **(c)** In XOR, truthful answer probabilities carry no information about the complementary relation; sharing the missing signal lets an exact solver recover the answer.

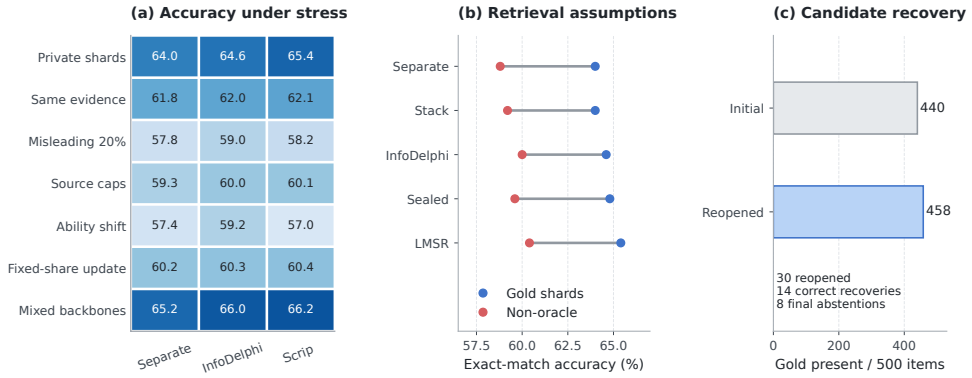

Figure 37: **The same vulnerabilities appear in the benchmark.** Qwen 2Wiki results. **(a)** A robustness matrix retains losses under misleading evidence and an ability shift. **(b)** Moving from controlled gold-shard access to non-oracle retrieval lowers all displayed methods. **(c)** Candidate diagnostics show what one reopening can recover, independently of final task correctness.

E Prompts and a Worked Episode

Prompts specify system permissions and use a common context across control methods: identical task text, menu metadata, candidate order, and currently available evidence. The system verifies the output schema and applies financial rules outside the model.

```
PRIVATE REPORT
You are {persona}. Answer {task} using only the evidence supplied.
The finalized candidate list is {candidate_ids_and_answers}.
OTHER means none of the listed answers. Assign probabilities summing
    to 1.
Record source IDs, one reason for and one against your leading
    candidate,
and what missing evidence would change your answer.
Return {probabilities, source_ids, concise_rationale}.
Do not infer that another agent has independent evidence.
```

```
ACQUISITION
Current balance: {balance}. Reserved collateral: {liability}.
Available services: {id, description_without_gold_relevance, price}.
```

```
Already read: {source_ids}. Current report: {probabilities}.
You may choose one service, an allowed bundle, or STOP.
Explain the possible usefulness and possible harm before seeing the
    content.
Do not assume expensive evidence is correct. Consider complements.
Return {action, service_ids, expected_change, reason}.
The controller will check affordability and charge real resource usage.
```

```
AMENDMENT (optional training arm only)
Given your persona and your last five training predictions with
    verified
outcomes, suggest a revised evidence-checking strategy. Do not memorize
answers, copy task-specific labels, or claim access to unavailable
    sources.
Return a persona with general instructions and a concise failure
    diagnosis.
```

E.1 Worked two-candidate episode

The ledger below follows a fictional two-candidate task from initial reports through settlement and normalization. Consider a fictional task: "Which of two sites, A or B, contains the archive?" Three

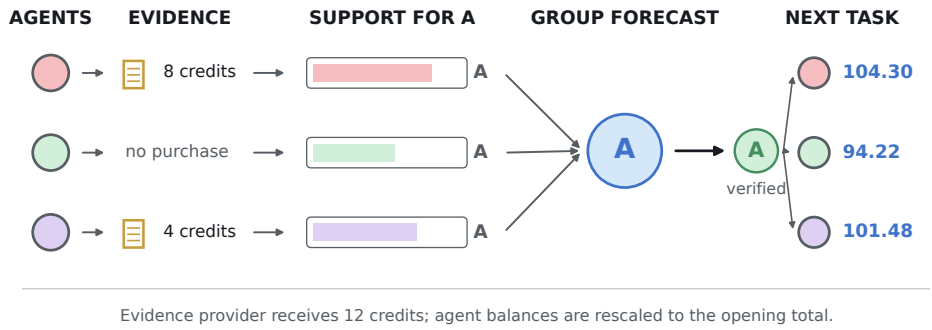

Figure 38: **A purchase changes both the current report and the next balance.** Accounting for the three-agent episode below; the evidence provider receives 12 credits and a logged common rescaling restores the opening total.

illustrative specialists have wealth $(100, 100, 100)$ and free-context reports for A of $(0.65, 0.55, 0.35)$. The initial equal pool is 0.5167. The three-agent construction keeps the ledger readable and serves as an accounting illustration alongside the 16-agent benchmark.

Agent 1 buys a corroborating source for 8 credits and revises to $q_1(A) = 0.8$. Agent 2 stops at 0.55. Agent 3 buys a location cross-reference for 4 credits and revises to 0.7. Their balances are $(92, 100, 96)$, stakes at $\kappa = .5$ are $(46, 50, 48)$, and total stake is 144. The sealed price is

$$p(A) = \frac{46(0.8) + 50(0.55) + 48(0.7)}{144} = \frac{97.9}{144} = 0.679861.$$

If the verified answer is A, the pre-normalization balances are $(100.129, 90.449, 97.422)$, summing to 288. The 12-credit evidence fee completes the original 300-credit ledger. Scaling by $300/288$ gives next-task balances $(104.301, 94.218, 101.481)$. The 12-credit house transfer is explicit, while the relative purchase penalty remains.

An equal pool over the same final reports gives 0.683333. A separate-budget pool with unchanged equal prediction weights gives the same value on this episode; differences emerge as costs and settlements alter future permissions. A debate solver pays to generate and read rationale messages, whose value depends on their content. If the sources carried complementary bits, a rationale could

convey the missing signal. This fictional episode illustrates the accounting, while measured traces establish benchmark performance.

E.2 Records for a real process case

A benchmark case study records task ID, candidate-generation stages, initial reports, service descriptions and prices, purchased source hashes, revised reports, every trade or stake, final probabilities, ground truth revealed only afterward, and the next-task wealth state. Debate and online-pool traces use the same item and evidence allocation. The fictional arithmetic example above illustrates the accounting format; measured traces supply the empirical comparison.

F Extended Literature

The preceding Related Work section identifies the direct experimental comparisons. This appendix places them in their longer intellectual context. The common theme is how dispersed knowledge, limited resources, and outcome feedback shape a collective decision.

Economic coordination. The motivating observation that knowledge is dispersed (Hayek, 1945) has a long computational lineage: agoric computation (Miller & Drexler, 1988), market-oriented programming (Wellman, 1993), bucket-brigade credit assignment (Holland, 1985; Schmidhuber, 1989), the society of mind (Minsky, 1988), and cooperative artificial economies (Baum, 1999; Baum & Durdanovic, 2000). These connections motivate resource accounting, while contemporary LLM populations require direct empirical evaluation of independent information acquisition. Mechanism design for LLMs and auction evaluation also study strategic behavior (Dütting et al., 2024; Chen et al., 2023a). Agent populations are evaluated here as software subject to allocation constraints and observable response policies.

Markets and reports. LMSR (Hanson, 2003; 2007), proper scoring (Brier, 1950; Savage, 1971; Gneiting & Raftery, 2007), and self-financed wagering (Lambert et al., 2008) support different mechanism designs. Interpretation of market prices requires assumptions about beliefs, utility, and constraints (Wolfers & Zitzewitz, 2006). Empirical successes of prediction markets (Wolfers & Zitzewitz, 2004; Berg et al., 2008; Arrow et al., 2008) coexist with strong deliberative forecasting teams (Atanasov et al., 2017; Mellers et al., 2014; Schoenegger et al., 2024). Dynamic aggregation depends on signal structure and strategic behavior (Ostrovsky, 2012; Chen & Waggoner, 2017). Field evidence that manipulation can persist (Rasooly & Rozzi, 2025) places manipulation resistance beyond what a collateral bound establishes.

Collective reasoning and failure. Debate was also proposed for scalable oversight (Irving et al., 2018; Brown-Cohen et al., 2023); market making has been discussed in that context (Hubinger, 2020). Practical collaboration spans software agents (Qian et al., 2024; Chen et al., 2023c), generative social agents (Park et al., 2023), and equilibrium generation/discrimination (Jacob et al., 2024). Controlled studies report limited or negative debate gains in some settings (Smit et al., 2024; Wang et al., 2024b); coordination failures (Cemri et al., 2025), sycophancy (Sharma et al., 2023), and social conformity (Zhang et al., 2024) make relevant diagnostic targets. Useful alternatives include relevance-aware routing, sparse communication, localized dispute resolution, and conditional rationale purchase. Oracle resolution is a separate task from forecasting; Kota (2026) study multi-agent resolution accuracy, so their scores belong in a separate resolution-accuracy report.

Confidence, provenance, and resource use. Verbalized confidence is useful but imperfect (Xiong et al., 2024; Tian et al., 2023; Kadavath et al., 2022); betting frames (Todasco, 2025) modify elicitation, while controller-enforced settlement changes weights. Dutch-book coherence (Andrews & Sarkar, 2026), source-level Sybil resistance (Bara, 2026), and propagation-aware uncertainty (Liu et al., 2026) diagnose different weaknesses. A shared hidden evidence source can create many confident reports without new information. Retrieval and acting (Lewis et al., 2020; Yao et al., 2023), query-adaptive inference (Snell et al., 2024), and cost-sensitive routing (Chen et al., 2023b) provide useful acquisition controls. Label-query budgets (Efroni et al., 2021) differ from evidence-feature budgets (Fainman et al., 2019); both differ from assigning prediction weight according to remaining cash.

G LIMITATIONS AND FAILURE CASES

The mechanism requires verifiable outcomes, a specified evidence interface, and auditable costs. The adaptive-seller stress test lowers accuracy in both budget arms, and seller incentives remain active under buyer-side accounting. The guarantees cover capital and maker loss under the stated fixed-seat protocol; collusion, unreliable labels, strategic evidence contents, and identity-dependent access require separate controls.

The measured gains arise when agents can choose among useful and redundant sources and the budget makes purchases consequential. On a fixed report trace, replay reduces sealed Scrip to a cost-adjusted pool; complementary signals require communication, and historical wealth can recover slowly when expertise reverses. Provenance-aware acquisition, task-dependent forgetting, explicit bundle reasoning, and stronger deliberation controls are natural next tests. Consequential deployments require calibration, deployment-volume cost accounting, and human review.