ARRIVAL-CONDITIONED VOTING POWER IN HETEROGENEOUS LANGUAGE MODEL ENSEMBLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) ensembles can reduce latency by voting before every response finishes. With heterogeneous models, completion order also determines whose responses count. We derive how this selection changes the response distribution and final decision, separating changes in model participation from selection within each model. Across two disjoint five-model panels spanning seven benchmarks, voting on the first three responses lowers selected-response accuracy by 5.09 and 11.84 percentage points. Changes in model participation explain most of both losses, and exhaustive subset controls distinguish arrival selection from the effect of using fewer votes. Changing delivery times alone realizes all 128 predicted decision changes with fixed responses. We then derive a completion-invariant rule: commit when every possible completion gives the same full-panel decision. This rule matches all 672 full-panel decisions while reducing mean latency by 40.83% and 22.49% on the two panels. Completion order thus changes effective voting weights, but some latency savings remain available without changing the full-panel decision.

1 INTRODUCTION

Parallel large language model (LLM) systems aggregate responses from samples, models, or agent roles (Wang et al., 2022; Jiang et al., 2023b; Li et al., 2024; Wang et al., 2024; Li et al., 2025), yet these responses rarely finish together. A latency-aware controller must choose between waiting for the full panel and deciding from early completions. With heterogeneous models, early aggregation changes panel membership and therefore the probability that each identity affects the decision. Figure 1 contrasts two stopping points on the same response sequence: after three completions, or once every possible unseen answer preserves the full-panel decision.

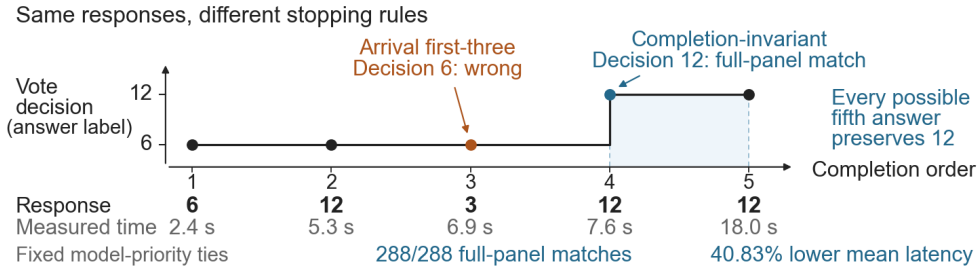

Figure 1: Same responses, different commit rules. The line shows the plurality decision after each completion, with ties resolved by fixed model priority. Vertical positions denote answer labels. Arrival first-three returns 6 incorrectly. At 7.6 s, every possible fifth answer preserves 12, allowing completion-invariant commit. Across 288 traces, this rule matches every full decision and reduces mean latency by 40.83%.

Adaptive inference makes this selection decision increasingly common. Aegean stops on stable semantic agreement (Ruan et al., 2025), while Semantic Quorum Assurance accepts outputs through weighted validation and veto rules (He & Yu, 2026). Learned orchestration removes unnecessary agent calls, and sequential evidence rules stop sampling as consensus develops (Shi et al., 2026;

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.

Medina, 2026). Scheduling and parallel execution also reduce end-to-end latency (Lin et al., 2024; Luo et al., 2025; Huang et al., 2026; Xu et al., 2026). Aggregate accuracy, however, does not reveal whether fast model identities systematically displace slow ones from the decision.

Equal plurality weights do not ensure equal participation. Although every model has one nominal vote, its probability of entering the first three depends on runtime. In both evaluated panels, the highest-accuracy model is almost absent from the first three.

We derive this effect directly from the joint law of responses and completion times. The selected-response law expresses how identity-specific inclusion propensities reweight semantic outputs. For fixed-size selection, an exact decomposition separates the resulting feature shift into identity reweighting and within-model selection. The joint decision law then retains response dependence and the nonlinear aggregator. We also derive a closed-form delay budget for reachable first-k decisions and a completion-invariant condition for the earliest partial history that guarantees the full-ensemble decision.

We test each link in this argument and repeat the analysis on a disjoint panel. In the reasoning panel, arrival first-three lowers selected-response accuracy by 5.09 points, of which identity reweighting accounts for 4.05 points, or 79.5%. Exhaustive random-three replay separates arrival selection from using fewer votes, and fixed-payload ingress replay realizes all 128 predicted reachable decision changes. A held-out panel with five different models and four non-mathematical benchmarks reproduces a larger 11.84-point selected-response loss under a parser frozen before generation. Completion-invariant commit matches every full decision in both panels while reducing mean latency by 40.83% and 22.49%.

Controllers can therefore prioritize early completions, preserve equal identity inclusion through random preselection, or preserve the full-panel decision through completion invariance. Our laws and certificate state what each choice guarantees.

2 Related Work

LLM ensembles combine independent samples, different models, or agent roles. Self-consistency aggregates sampled reasoning paths, and LLM-Blender ranks and fuses outputs (Wang et al., 2022; Jiang et al., 2023b). Mixture-of-Agents and More Agents combine multiple agent responses, while Sparse Mixture-of-Agents (SMoA) reduces interaction through response selection and early stopping (Wang et al., 2024; Li et al., 2024; 2025). Inverse Surprising Popularity (ISP) uses first- and second-order response information for training-free aggregation under model heterogeneity (Ai et al., 2025). Semantic Quorum Assurance assigns explicit validator weights and vetoes (He & Yu, 2026), while Aegean formalizes semantic agreement and early quorum (Ruan et al., 2025). These methods specify aggregation or stopping rules. We isolate how completion order induces identity weights even when declared votes are equal.

Adaptive inference chooses a model, depth, schedule, or future call from partial evidence. Confident Adaptive Language Modeling (CALM) exits on intermediate confidence, while FrugalGPT and RouteLLM route under quality and cost objectives (Schuster et al., 2022; Chen et al., 2023; Ong et al., 2024). Pairwise learning-to-rank scheduling and NexusSched predict service behavior for scheduling, routing, and batching (Tao et al., 2025; Zhang et al., 2025). Agent-serving systems optimize critical paths and redundant work (Lin et al., 2024; Luo et al., 2025; Huang et al., 2026; Li et al., 2026; Xu et al., 2026). Certaindex allocates reasoning budgets from answer stability (Fu et al., 2025); learned orchestration selects agent calls, and sequential evidence rules control when sampling stops (Shi et al., 2026; Medina, 2026). These methods decide where or how long to compute. Our analysis asks which identities enter the aggregate and certifies when all admissible unseen terminal responses preserve the full-panel decision.

Related mechanisms appear in distributed systems and unequal-probability sampling. Hedged requests reduce tail latency by accepting early replicas (Dean & Barroso, 2013). Fast-worker distributed optimization similarly trades completion time against the composition and staleness of accepted updates (Hanna et al., 2022; Alahyane et al., 2026). Unequal-probability sampling formalizes how inclusion propensities change population estimates (Horvitz & Thompson, 1952). These works establish the importance of inclusion mechanisms, but they do not model heterogeneous LLM responses as semantic marks whose selected joint configuration enters a nonlinear aggregator.

LLM timing has also been studied as an information and resource channel. Remote timing can reveal properties of efficient inference, and prompt-cache audits can infer hidden service behavior (Carlini & Nasr, 2024; Gu et al., 2025). Load Hijack uses poisoned router weights and trigger inputs to create stragglers in mixture-of-experts inference (Zhang et al., 2026). Combined software and hardware attacks can amplify security impact in compound artificial intelligence (AI) systems (Banerjee et al., 2026). Runtime logic has also been proposed for distributed LLM-agent workflows (Bollig, 2026). We study a complementary integrity effect: with an arrival-conditioned aggregator, nonnegative delivery delay can change the selected semantic evidence and the final decision while leaving every response unchanged.

3 Arrival-Conditioned Selection

3.1 Time, responses, and selection

We model one request with context $X = x$ and n models. Model i produces a terminal semantic mark $Y_i \in \mathcal{Y}$ at completion time $L_i \in \mathbb{R}_+$. A mark can be an answer label, a structured judgment, or an explicit abstention. The marked arrival process is

$$N_x(dt, dy, di) = \sum_{i=1}^{n} \delta_{(L_i, Y_i, i)}(dt, dy, di). \tag{1}$$

The visible history generates the filtration $\mathcal{F}_t = \sigma(X, U, N_x([0, s] \times B \times I) : s \le t)$ over measurable mark sets B and identity sets I, where U records any policy randomness sampled before execution. A streaming policy selects a stopping time τ adapted to $\mathcal{F}_t$, a selected identity set S, and a final decision $D = A(Y_S)$ under aggregator A. Static first-k uses the first time at which k terminal marks are visible. This formulation makes completion time part of the selection rule without assuming a Transformer architecture or independent model outputs. The same formulation applies after averaging over a request distribution. Our empirical estimates use the evaluated benchmark mixture as that distribution.

3.2 Selected-response law and effective voting power

Arrival-conditioned voting power is determined by inclusion probability. For a fixed-size policy with $|S| = k$, define $I_i = \mathbf{1}\{i \in S\}$ and

$$\pi_i(y, \ell \mid x) = \Pr(I_i = 1 \mid Y_i = y, L_i = \ell, X = x). \tag{2}$$

Let Q_x be the law of a response drawn uniformly from the selected set, and let P_i be the joint marginal law of (Y_i, L_i) given x. Conditioning on each response-time pair gives

$$\boxed{Q_x(B) = \frac{1}{k} \sum_{i=1}^{n} \int_{B \times \mathbb{R}_+} \pi_i(y, \ell \mid x) P_i(dy, d\ell \mid x).} \tag{3}$$

The nominal identity weight is $1/n$, whereas the effective identity weight is p_i/k with $p_i = \mathbb{E}[I_i \mid x]$. Equation 3 requires no independence, exchangeability, or continuous-time assumption. It states that a runtime inclusion policy reweights the semantic mark law through its own propensity.

The main diagnostic separates identity reweighting from selection within an identity. For any integrable response feature $\phi_i = \phi(Y_i)$, let $a_i = \mathbb{E}[\phi_i \mid x]$ and $c_i = \mathrm{Cov}(I_i, \phi_i \mid x)$. Relative to the nominal all-model mixture, the selected-response feature shift is

$$\boxed{\Delta_\phi(x) = \sum_{i=1}^{n} \left(\frac{p_i}{k} - \frac{1}{n} \right) a_i + \frac{1}{k} \sum_{i=1}^{n} c_i.} \tag{4}$$

The first sum is the identity-mixture term. It measures alignment between inclusion rates and feature means across model identities. The second sum is the within-model term. It captures dependence between a model's feature value and its inclusion event. The equality follows from $\mathbb{E}[I_i \phi_i] = p_i a_i + c_i$ and remains exact for empirical averages. With ϕ_i equal to answer correctness, the two terms show whether selected-response accuracy changes because the policy chooses different models or different responses from the same model.

3.3 From response weights to final decisions

First-order response weights do not determine a nonlinear aggregate decision. Let $\Pi_{\Gamma,x}(s \mid y, \ell)$ be the subset kernel induced by policy Γ after integrating over its explicit randomness, and let P_x be the joint law of all marks and completion times. For any decision set C,

$$\Pr_{\Gamma}(D \in C \mid x) = \sum_{s \subseteq [n]} \int \mathbf{1}\{A(y_s) \in C\}\, \Pi_{\Gamma,x}(s \mid y, \ell) P_x(dy, d\ell). \tag{5}$$

This decision law retains response dependence and the aggregator. It also motivates an exact control: uniformly preselecting one of the $\binom{n}{k}$ identity subsets before generation sets every $p_i = k/n$ and makes the selected-response identity term zero. Its decision can still differ from full-n aggregation because A receives fewer marks. Comparing arrival first-k, uniform preselection, and full aggregation therefore separates arrival selection from the fewer-vote effect.

3.4 Delay-only intervention and exact early commit

A delay-only intervention changes effective completion times to $\widetilde{L}_i = L_i + \delta_i$ with $\delta_i \geq 0$ and leaves Y_i fixed. For a realized trace and target k-subset S, define $t_S = \max_{j \in S} L_j$. Under static first-k with a fixed mark-independent tie rule, the infimum total delay needed to make S arrive first is

$$\boxed{B_{\inf}(S) = \sum_{i \notin S} [t_S - L_i]_+,} \tag{6}$$

where $[u]_+ = \max(u, 0)$. Minimizing this value over subsets whose aggregate differs from the natural first-k decision gives a per-trace decision-flip margin. The margin is computed from a completed trace and gives the minimum required delay under the stated intervention. Finite reachability also requires an alternative subset that yields another decision.

Exact early commit uses a different principle. Given a visible history h_t, let $\mathfrak{C}(h_t)$ contain every admissible completion of the unseen marks, and define

$$\mathcal{D}(h_t) = \{A(h_t \oplus c) : c \in \mathfrak{C}(h_t)\}. \tag{7}$$

The policy commits exactly when $|\mathcal{D}(h_t)| = 1$. This condition is sufficient because the realized continuation belongs to $\mathfrak{C}(h_t)$. It is also necessary for any deterministic policy that must agree with every admissible full completion: two completions with different final decisions share the same visible history. For plurality with the fixed identity-priority tie rule used below, Proposition 6 reduces this condition to an exact $O(n + m)$ certificate, where m is the number of visible labels. Appendix B gives complete proofs, the static first-k inclusion formula, and the non-identifiability of the nominal mixture from selected-only traces.

4 Experimental Design

The reasoning panel identifies the mechanism on structured numeric answers. It contains Qwen2.5 0.5B and 1.5B, Qwen3 0.6B, Gemma 2 2B, and Llama 3.1 8B (Qwen Team, 2024; 2025; Gemma Team, 2024; Llama Team, 2024). Here, B denotes one billion parameters. Each model answers 48 Grade School Math 8K (GSM8K), 48 Simple Variations on Arithmetic Math Word Problems (SVAMP), and 48 MATH benchmark problems (Cobbe et al., 2021; Patel et al., 2021; Hendrycks et al., 2021b). Two deterministic executions per model-task pair produce 144 task clusters, 288 traces, and 1,440 generations. The prompt requests one final numeric answer, and unresolved endpoints map to an abstention mark.

The held-out panel tests the same mechanism with different models, tasks, and answer format. It contains OLMo 2 1B, SmolLM2 1.7B, Granite 3.3 2B, Phi-3.5 Mini, and Mistral 7B (Team OLMo et al., 2024; Ben Allal et al., 2025; IBM Granite Team, 2025; Abdin et al., 2024; Jiang et al., 2023a). Each model answers 48 held-out tasks from the Artificial Intelligence 2 Reasoning Challenge (ARC-Challenge), Massive Multitask Language Understanding (MMLU), TruthfulQA, and HellaSwag (Clark et al., 2018; Hendrycks et al., 2021a; Lin et al., 2022; Zellers et al., 2019). A multiple-choice parser was frozen before held-out generation. Two executions produce 192 task clusters, 384 traces,

and 1,920 generations. Appendix E specifies the disjoint task selection, parser, and unresolved-answer bounds.

Both panels use complete shadow traces. We measure each model in isolation and replay the five completions as virtual workers that start together. This protocol preserves every potential response under each policy, which is required to estimate the nominal mixture and replay counterfactual subsets. Service time is measured for each completed sequence under the batched decoder specified in Appendix C. The main estimand averages the two repeated executions within each task before inference across tasks.

Full-five plurality is the reference. Arrival first-three selects the earliest three marks. Uniform random-three averages exactly over all ten identity subsets chosen before seeing outputs or times. Completion-invariant plurality applies Equation 7. Quorum-3 else full commits when three matching labels arrive and otherwise waits for all five. A fixed identity order resolves ties and is chosen before outcomes. We call a policy decision-equivalent when it matches full-five plurality on every trace. We report selected-response accuracy, defined as the mean correctness of selected marks, together with decision accuracy, inclusion propensity, full-decision disagreement, and relative commit-latency reduction. Accuracy differences use percentage points (pp). The two panels use 95% confidence intervals (CIs) from 10,000 task-cluster bootstrap replicates. Finite sensitivity checks enumerate $k = 1, \ldots, 5$ and all $5!$ fixed identity tie orders. Appendix C specifies data selection, parsing, tie handling, and estimands.

We replay all 288 traces through a controlled Transmission Control Protocol (TCP) ingress process. For each of the 128 traces with a reachable alternative first-three decision, Equation 6 supplies nonnegative delivery delays for a target subset. The replay preserves all response payloads and sends them in the order specified by the delayed schedule. We verify received order, selected subset, aggregate decision, and payload identity for each replay.

5 Results

5.1 Model participation under arrival selection

Arrival first-three creates a strongly nonuniform identity mixture. Figure 2a compares each model's inclusion propensity with its answer accuracy. Llama 3.1 8B has the highest accuracy at 59.72%, yet it never appears among the first three. The two Qwen2.5 models appear in 85.07% and 86.11% of selected sets. The selected-response accuracy is 41.44%, compared with 46.53% for the nominal five-model mixture.

The decomposition attributes most of this loss to model identity. Arrival first-three reduces selected-response accuracy by 5.09 points (95% CI, 2.73 to 7.45). Identity reweighting contributes 4.05 points, or 79.5% of the observed loss, whereas within-model selection contributes 1.04 points and its CI includes zero. The two terms reconstruct the observed change to numerical precision.

5.2 Separating vote count from arrival selection

Table 1 separates the cost of using three votes from the cost of selecting them by arrival. As required by Equation 4, averaging all ten preselected subsets produces exactly zero selected-response shift. Decision accuracy still falls from 56.94% to 50.83%, a 6.11-point loss (95% CI, 3.06 to 9.51), because three-vote plurality can differ from five-vote plurality. Arrival first-three falls further to 42.71%. Relative to uniform random-three, this contrast isolates an additional 8.12-point loss from arrival-conditioned subset selection (95% CI, 3.68 to 12.71).

Figure 3 makes the policy distinction visible. Random preselection saves less latency because its chosen subset can include slow identities. Arrival first-three saves 67.14% on average but lowers decision accuracy by 14.24 points. Completion-invariant commit and quorum-3 else full stay on the exact-equivalence line. Completion invariance gives the larger decision-equivalent reduction, 40.83%, because it can commit under any partial vote pattern whose unseen completions cannot change the result.

Table 1: Overall policy comparison on 288 panel traces. Accuracy changes use full-five plurality as reference. Latency entries report the mean followed by its 95% task-cluster bootstrap CI in parentheses. Decision-equivalent policies agree with the full decision on every trace.

Policy	Decision accuracy (%)	Δ accuracy (pp)	Latency reduction (%)
Full five	56.94	0.00	0.00
Uniform random-three	50.83	−6.11	24.16 (23.32–25.00)
Arrival first-three	42.71	−14.24	67.14 (65.36–68.90)
Quorum-3 else full	56.94	0.00 exact	35.57 (30.33–40.76)
Completion-invariant	56.94	0.00 exact	**40.83** (35.93–45.77)

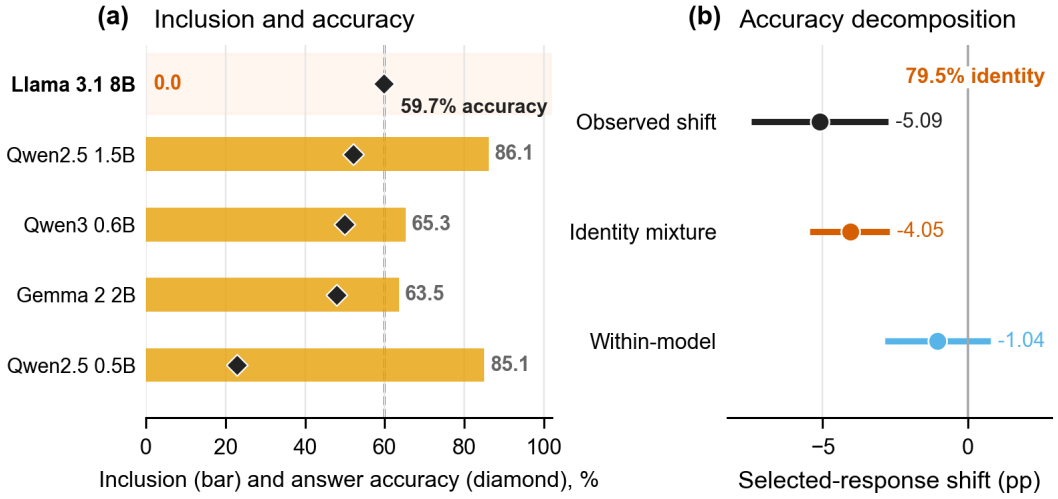

Figure 2: Arrival-conditioned identity reweighting explains most of the selected-response loss. (a) Bars show first-three inclusion, diamonds show base answer accuracy, and the dashed line marks equal 60% inclusion. The highest-accuracy model has zero inclusion. (b) Equation 4 reconstructs the observed 5.09-point loss exactly. Lines show 95% task-cluster bootstrap intervals.

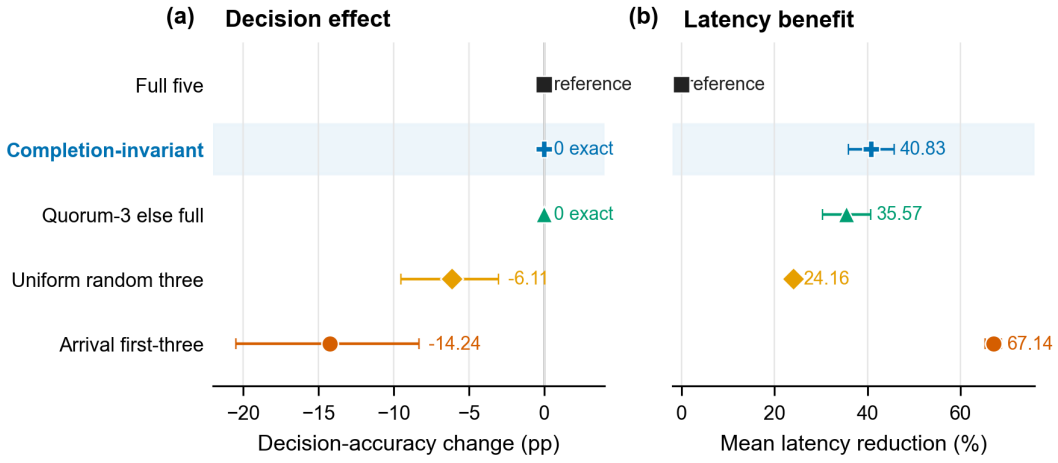

Figure 3: Completion-invariant commit provides the largest latency reduction among decision-equivalent policies. The aligned panels report decision-accuracy change and mean latency reduction for the same five policies. Completion-invariant and quorum-3 else full match every full decision exactly. Horizontal lines show 95% task-cluster bootstrap intervals where applicable.

The held-out panel confirms identity-driven selection with different models, benchmarks, and answer format. Arrival first-three lowers selected-response accuracy by 11.84 points (95% CI, 9.86 to 13.82), and the loss remains at least 10.17 points under every assignment to unresolved answers. Identity reweighting contributes a 12.19-point decrease, while within-model selection is near zero. This pattern follows model exposure: OLMo 2 and SmolLM2 enter every first-three set, whereas Phi-3.5 Mini combines 79.17% base accuracy with 2.08% inclusion.

Response-preserving latency mappings provide a direct mechanism check. Equalizing identity medians changes the native 11.84-point loss to a 1.09-point gain whose interval includes zero. Reassigning each within-trace time multiset in reverse identity-rank order yields a 14.55-point gain. Equation 4 reconstructs both changes, with identity reweighting dominating the reversal. Table 2 summarizes the native recurrence across panels. Appendix Figure 5 and Tables 17–20 give the complete results.

Table 2: The held-out panel reproduces identity-driven selected-response change with different models and tasks. Decision-equivalent reduction is the mean latency reduction of completion-invariant commit, which has zero full-decision mismatches in both panels.

Panel	Tasks / traces	Selected shift (pp)	Identity term (pp)	Decision-equivalent reduction (%)
Reasoning	144 / 288	-5.09	-4.05	40.83
Held-out choice	192 / 384	-11.84	-12.19	22.49

The held-out final decisions show the same separation between vote count and arrival order. Arrival first-three lowers decision accuracy by 22.40 points and disagrees with full-five plurality on 29.69% of traces. Exhaustive uniform random-three shows that 15.16 points of this loss arise beyond the use of three votes. Completion-invariant commit matches all 384 full decisions and reduces mean latency by 22.49%. Appendix E reports the corresponding CIs and complete policy table.

Finite enumeration on the reasoning panel shows that this conclusion does not depend on one value of k or one tie order. The selected-response shift approaches zero as k increases, and all 120 fixed identity tie orders retain negative arrival-versus-full and arrival-versus-random-three decision shifts (Appendix Tables 9 and 10).

Completion invariance saves more time because it accepts more decisive vote patterns than a three-label quorum. In Figure 1, the fourth response changes the partial decision from 6 to 12. The visible labels $6, 12, 3, 12$ then fix the full winner: an unseen 12 strengthens the lead, 6 or 3 gives 12 a fixed-priority 2–2 win, and a new label receives only one vote. Across 288 traces, completion invariance commits before the fifth response 192 times, compared with 166 for quorum-3 else full (Appendix Table 8).

The plurality certificate also has linear evaluation cost. Proposition 6 reduces the test to at most one continuation per visible label plus one fresh-label continuation. With cached vote counts, the certificate takes $O(n + m)$ time for m visible labels. An independent implementation agrees with direct continuation enumeration on all 5,286 canonical states through $n = 7$ and 10,000 additional states for $n = 8, \ldots, 16$ (Appendix B).

5.3 Disagreement at plurality boundaries

Most natural disagreements occur where subset membership can change a plurality outcome. Arrival first-three and full-five disagree on 66 of 288 traces. Sixty-two of these disagreements lie in the 96 traces with a plurality tie in either the first-three or full-five vote, while only four occur among 192 traces without either tie. All 66 disagreements occur in traces containing at least two distinct terminal marks. The 64 traces whose five marks are identical have zero selected-response and decision shift. Figure 4a shows this localization. The result explains why mean response quality alone cannot predict the decision effect: the joint vote configuration determines whether a membership change crosses an aggregation boundary.

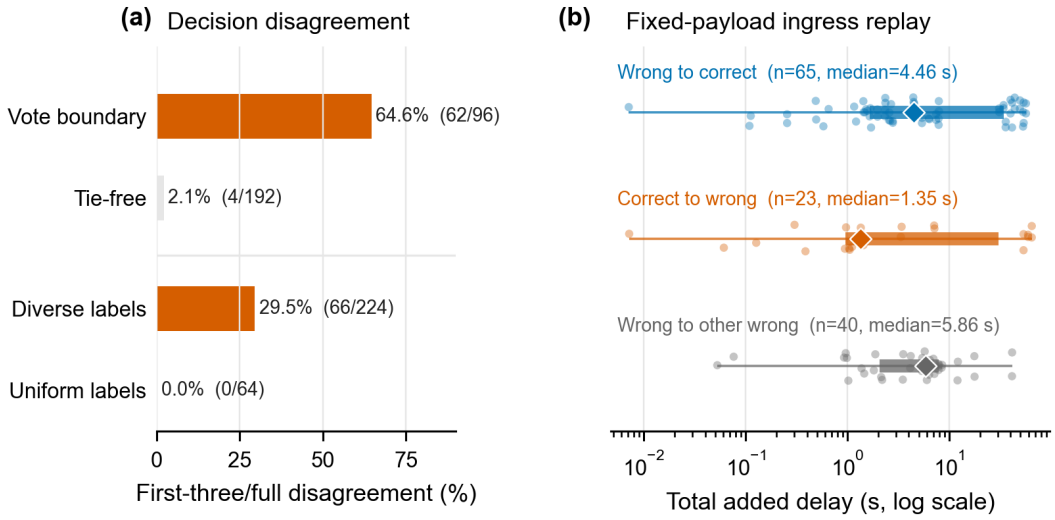

Figure 4: Decision changes localize near aggregation boundaries and remain reachable by delay alone. (a) Bars report first-three/full disagreement rates within two overlapping diagnostics. (b) Nonnegative delivery delay realizes all 128 predicted reachable target decisions with identical payloads. Points show verified interventions. Thick and thin lines show interquartile and full ranges, and diamonds mark medians.

5.4 Fixed-response delay interventions

The controlled ingress replay realizes every target predicted by Equation 6. All 288 natural traces reproduce their baseline arrival order and decision. Among the 128 traces with an alternative reachable decision, all 128 delayed replays preserve payload bytes, realize the target first-three subset, and produce the target decision. This set includes 23 correct-to-wrong changes with median total added delay 1,351 ms, 65 wrong-to-correct changes with median 4,461 ms, and 40 changes between two wrong answers. These replays verify that the controller changes its decision under the constructed delivery schedules while response payloads remain fixed.

Endpoint sensitivities support the same mechanism. Requiring every model to finish at a valid answer endpoint leaves 122 traces and yields an 11.15-point selected-response loss and a 25.41-point decision loss. Thus, invalid endpoints do not create the aggregate arrival effect. Exact outputs match across the two greedy executions for every generation, and 95.14% of panels preserve the complete arrival rank. Appendix D reports the full sensitivity tables and deterministically selected examples.

Shared execution also permits decision-preserving early commit. On one graphics processing unit (GPU), the reasoning panel processes six 24-task workloads under paired isolated and shared conditions, producing identical output tokens. Completion invariance matches all 288 shared-execution full decisions and reduces replayed commit time by 12.91% (95% workload-cluster CI, 10.90 to 14.93). The reduction is positive in every workload and remains 11.70% on tasks excluded from the deployment pilot. Arrival first-three reaches 48.26% accuracy, compared with 59.03% for full five. Appendix G specifies the paired timing protocol and reports all policies and task groups.

6 Discussion

Effective voting power is a deployment property. The aggregator, model panel, and runtime policy jointly determine which evidence affects a decision. Equal plurality weights coexist with unequal identity weights whenever completion order changes inclusion propensity, extending semantic ensemble work on response diversity and explicit validator weights (Wang et al., 2024; Li et al., 2025; He & Yu, 2026). Both panels place accurate identities late, and response-preserving latency remapping moves the identity term. Evaluations should therefore report identity-level inclusion with answer quality. Complete shadow traces expose every potential response, while random preselection guarantees equal identity inclusion. Inverse weighting cannot recover an identity with zero inclusion

(Appendix B). Since final decisions also depend on the joint vote configuration, propensity diagnostics should be paired with disagreement and decision accuracy.

Latency policies should state what they preserve. Early completion can reduce critical-path latency (Ruan et al., 2025; Shi et al., 2026; Lin et al., 2024; Luo et al., 2025; Fu et al., 2025), but each rule preserves a different property. Uniform preselection preserves equal identity inclusion. Arrival first-k prioritizes the earliest completions and gives the largest measured latency reduction. Completion-invariant commit preserves the full decision, matches all 672 model-isolated full-panel decisions, and requires no calibration.

Identity reweighting and decision equivalence answer different design questions. Equation 4 measures how a fixed-size selected set changes response composition, whereas Equation 7 asks whether the visible labels already determine the full aggregate. Equal inclusion removes the identity-mixture term but need not preserve the plurality decision because the joint label configuration still enters a nonlinear rule. Completion invariance instead conditions on the realized vote state and preserves the full decision without requiring equal exposure or an accuracy model. Systems seeking equal identity participation should report inclusion propensities. Systems seeking the full-panel decision should provide a completion certificate.

Exact early commit separates a general condition from an efficient implementation. Equation 7 applies to any deterministic aggregator whose admissible completion set contains the realized continuation. Proposition 6 gives the plurality-specific implementation with at most $m + 1$ extremal completions and $O(n + m)$ time. An independent implementation agrees with direct enumeration through panel size 16. The guarantee is relative to a fixed parser and aggregator. Ranking, score fusion, and open-ended aggregation require their own completion representation and efficient certificate.

Timing can affect semantic integrity. Prior work establishes timing leakage, hidden cache behavior, and resource manipulation in LLM inference (Carlini & Nasr, 2024; Gu et al., 2025; Zhang et al., 2026; Banerjee et al., 2026). Our ingress replay shows a direct decision effect: delivery delay changes which fixed responses are counted. Equation 6 gives the exact static first-k margin, and the replay verifies the corresponding controller intervention. Completion-invariant commit prevents delivery order from changing the fixed full-aggregation target because every admissible unseen completion yields the same decision at commit time.

The experiments show why aggregate speed-accuracy curves are insufficient. The evidence spans ten models and seven benchmarks. Paired shared execution adds a deployment check on the reasoning panel, while complete traces distinguish identity selection from changes in the joint vote configuration. The exact laws allow dependent responses and completion times. Our measured commit reductions retain all responses for replay; live request queues and cancellation remain outside these measurements.

7 Conclusion

Arrival-based selection assigns effective voting power through completion order. The exact decomposition shows whether a selected-response shift comes from model identity or within-model response variation, and the joint decision law explains why this first-order shift must be evaluated together with aggregation boundaries. The identity-driven shift appears in two independent five-model panels spanning seven benchmarks, while exhaustive subset controls separate arrival conditioning from using fewer votes. Delay-only replay confirms that fixed response payloads can yield different decisions when delivery order changes. Completion-invariant commit provides a constructive alternative: it matches all 672 full decisions while retaining 40.83% and 22.49% mean latency reductions across the two panels. These results make identity participation an explicit part of latency-aware ensemble design.

AI use statement

Generative AI tools supported mathematical formalization, experimental code implementation and checking, and manuscript polishing. The authors developed the theoretical framework and experimental design, interpreted the results, and verified the citations, equations, code, figures, insights, claims, and prose. The authors take responsibility for all content.

Ethics statement

This study uses public models and benchmarks and involves no human subjects or private data. The delay-only analysis identifies an integrity risk under stated intervention assumptions. We report both outcome directions and an exact decision-preserving rule. We do not evaluate an adaptive attack on a public service.

Reproducibility statement

Appendices C, E, and G specify the protocols, Appendix B gives complete proofs, and Appendices D, E, and G report the results. The Supplementary Material provides the complete experimental code and results, including raw generations, policy replays, ingress interventions, statistical analyses, and figure-generation code.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. URL https://arxiv.org/abs/2404.14219.

Rui Ai, Yuqi Pan, David Simchi-Levi, Milind Tambe, and Haifeng Xu. Beyond majority voting: LLM aggregation by leveraging higher-order information. *arXiv preprint arXiv:2510.01499*, 2025. URL https://arxiv.org/abs/2510.01499.

Abdelkrim Alahyane, Céline Comte, and Matthieu Jonckheere. Optimization trade-offs in asynchronous federated learning: A stochastic networks approach. *arXiv preprint arXiv:2603.26231*, 2026. URL https://arxiv.org/abs/2603.26231.

Sarbartha Banerjee, Prateek Sahu, Anjo Vahldiek-Oberwagner, Jose Sanchez Vicarte, and Mohit Tiwari. Cascade: Composing software-hardware attack gadgets for adversarial threat amplification in compound AI systems. *arXiv preprint arXiv:2603.12023*, 2026. URL https://arxiv.org/abs/2603.12023.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big: Data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025. URL https://arxiv.org/abs/2502.02737.

Benedikt Bollig. Causal past logic for runtime verification of distributed LLM agent workflows. *arXiv preprint arXiv:2605.20923*, 2026. URL https://arxiv.org/abs/2605.20923.

Nicholas Carlini and Milad Nasr. Remote timing attacks on efficient language model inference. *arXiv preprint arXiv:2410.17175*, 2024. URL https://arxiv.org/abs/2410.17175.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023. URL https://arxiv.org/abs/2305.05176.

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

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

Jeffrey Dean and Luiz André Barroso. The tail at scale. *Communications of the ACM*, 56(2):74–80, 2013. doi: 10.1145/2408776.2408794.

Yichao Fu, Junda Chen, Siqi Zhu, Zheyu Fu, Zhongdongming Dai, Yonghao Zhuang, Yian Ma, Aurick Qiao, Tajana Rosing, Ion Stoica, and Hao Zhang. Efficiently scaling LLM reasoning programs with Certaindex. In *Advances in Neural Information Processing Systems 38*, pp. 157566–157604, 2025. doi: 10.52202/085713-4751. URL https://doi.org/10.52202/085713-4751.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Chenchen Gu, Xiang Lisa Li, Rohith Kuditipudi, Percy Liang, and Tatsunori Hashimoto. Auditing prompt caching in language model APIs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20477–20496. PMLR, 2025. URL https://proceedings.mlr.press/v267/gu25b.html.

Serge Kas Hanna, Rawad Bitar, Parimal Parag, Venkat Dasari, and Salim El Rouayheb. Adaptive stochastic gradient descent for fast and communication-efficient distributed learning. *arXiv preprint arXiv:2208.03134*, 2022. URL https://arxiv.org/abs/2208.03134.

Jun He and Deying Yu. Semantic quorum assurance: Collective certification for non-deterministic AI infrastructure. *arXiv preprint arXiv:2606.08021*, 2026. URL https://arxiv.org/abs/2606.08021.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021b. URL https://arxiv.org/abs/2103.03874.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Yeqi Huang, Yanwei Ye, Guomin Chen, Wenhao Su, Bin Gong, Jialian Li, Zhan Lu, Yangshen Deng, Xuan Sun, Le Xu, and Luo Mai. SwarmX: Agentic scheduling for low-latency agentic systems. *arXiv preprint arXiv:2606.21401*, 2026. URL https://arxiv.org/abs/2606.21401.

IBM Granite Team. Granite 3.3 2B Instruct model card. Hugging Face model repository, 2025. URL https://huggingface.co/ibm-granite/granite-3.3-2b-instruct.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a. URL https://arxiv.org/abs/2310.06825.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 14165–14178. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.acl-long.792.

Dawei Li, Zhen Tan, Peijia Qian, Yifan Li, Kumar Chaudhary, Lijie Hu, and Jiayi Shen. SMoA: Improving multi-agent large language models with sparse mixture-of-agents. In *Advances in Knowledge Discovery and Data Mining*, pp. 54–65. Springer Nature Singapore, 2025. doi: 10.1007/978-981-96-8180-8_5.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *arXiv preprint arXiv:2402.05120*, 2024. URL https://arxiv.org/abs/2402.05120.

Yu Li, Qinyuan Ye, Prafulla Kumar Choubey, Jiaxin Zhang, and Chien-Sheng Wu. Speculate with memory: Lossless acceleration for LLM agents. *arXiv preprint arXiv:2607.12236*, 2026. URL https://arxiv.org/abs/2607.12236.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Chaofan Lin, Zhenhua Han, Chengruidong Zhang, Yuqing Yang, Fan Yang, Chen Chen, and Lili Qiu. Parrot: Efficient serving of LLM-based applications with semantic variable. In *18th USENIX Symposium on Operating Systems Design and Implementation*, pp. 929–945. USENIX Association, 2024. URL https://www.usenix.org/conference/osdi24/presentation/lin-chaofan.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.229.

Llama Team. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Michael Luo, Xiaoxiang Shi, Colin Cai, Tianjun Zhang, Justin Wong, Yichuan Wang, Chi Wang, Yanping Huang, Zhifeng Chen, Joseph E. Gonzalez, and Ion Stoica. Autellix: An efficient serving engine for LLM agents as general programs. *arXiv preprint arXiv:2502.13965*, 2025. URL https://arxiv.org/abs/2502.13965.

Roberto E. Medina. Adaptive consensus in LLM ensembles via sequential evidence accumulation: Automatic budget identification and calibrated commit signals. *arXiv preprint arXiv:2605.04236*, 2026. URL https://arxiv.org/abs/2605.04236.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. *arXiv preprint arXiv:2406.18665*, 2024. URL https://arxiv.org/abs/2406.18665.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.168.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Chaoyi Ruan, Yiliang Wang, Ziji Shi, and Jialin Li. Reaching agreement among reasoning LLM agents. *arXiv preprint arXiv:2512.20184*, 2025. URL https://arxiv.org/abs/2512.20184.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. *arXiv preprint arXiv:2207.07061*, 2022. URL https://arxiv.org/abs/2207.07061.

Xi Shi, Mengxin Zheng, and Qian Lou. Learning latency-aware orchestration for multi-agent systems. *arXiv preprint arXiv:2601.10560*, 2026. URL https://arxiv.org/abs/2601.10560.

Yiheng Tao, Yihe Zhang, Matthew Dearing, Xin Wang, Yuping Fan, Michael E. Papka, and Zhiling Lan. Ranking before serving: Low-latency LLM serving via pairwise learning-to-rank. *arXiv preprint arXiv:2510.03243*, 2025. URL https://arxiv.org/abs/2510.03243.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024. URL https://arxiv.org/abs/2501.00656.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024. URL https://arxiv.org/abs/2406.04692.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. URL https://arxiv.org/abs/2203.11171.

Zihan Xu, Haolin Tian, and Hai Jiang. A two-tier perspective on inference-time parallelism in multi-agent LLM systems. *arXiv preprint arXiv:2608.05791*, 2026. URL https://arxiv.org/abs/2608.05791.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1472.

Rui Zhang, Wenbo Jiang, Hongwei Li, Zihan Wang, Rui Zhang, Chaoshun Zuo, Jianfei Sun, and Guowen Xu. Trigger the straggler: Load hijack on mixture-of-experts LLMs. *arXiv preprint arXiv:2608.10614*, 2026. URL https://arxiv.org/abs/2608.10614.

Yue Zhang, Yuansheng Chen, Xuan Mo, Alex Xi, Jialun Li, and Weigang Wu. A predictive and synergistic two-layer scheduling framework for LLM serving. *arXiv preprint arXiv:2509.23384*, 2025. URL https://arxiv.org/abs/2509.23384.

A Appendix roadmap

The appendix mirrors the main argument. It begins with theory and the reasoning-panel protocol, then reports complete results, held-out confirmation, the ingress intervention, shared execution, and reproduction details. The table below gives the section map.

Appendix	Contents
B	Complete mathematical statements and proofs
C	Models, tasks, decoding, marks, timing, policies, and inference
D	Full quantitative results and sensitivity analyses
E	Held-out models, tasks, parser, bounds, policies, and remappings
F	Payload-fixed delay intervention and verification
G	Paired shared execution, workloads, policies, and subgroup results
H	Artifact inventory and deterministic reproduction steps

B Complete theory and proofs

B.1 General setting

Fix a deployment context $X = x$. We allow the vector $(Y_{1:n}, L_{1:n})$ to have an arbitrary joint law. Models may be heterogeneous, marks and completion times may be dependent, and model outputs may share latent randomness. A policy observes the filtration generated by Equation 1 and returns a selected set S. Unless stated otherwise, $|S| = k$ almost surely. We write $I_i = \mathbf{1}\{i \in S\}$. Every result below also holds for an unconditional run law obtained by integrating over a specified distribution of X. The empirical analysis uses the evaluated task mixture and recomputes uncertainty by resampling task clusters.

The nominal all-model mark mixture is

$$P_x^0(B) = \frac{1}{n}\sum_{i=1}^{n} \Pr(Y_i \in B \mid X = x). \tag{8}$$

To define the selected-response law, first sample one run and then draw one element uniformly from its selected set. For fixed k, this is equivalent to count weighting because every run contributes the same number of selected items.

B.2 Selected-response law

Theorem 1 (Selected-response propensity law)**.** *For every measurable $B \subseteq \mathcal{Y}$,*

$$Q_x(B) = \frac{1}{k}\sum_{i=1}^{n} \int_{B\times\mathbb{R}_+} \pi_i(y, \ell \mid x) P_i(dy, d\ell \mid x), \tag{9}$$

where $\pi_i(y, \ell \mid x) = \Pr(I_i = 1 \mid Y_i = y, L_i = \ell, X = x)$.

Proof. Conditional on the selected set, uniform sampling gives

$$Q_x(B) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{E}[I_i \mathbf{1}\{Y_i \in B\} \mid X = x]. \tag{10}$$

Conditioning each summand on (Y_i, L_i) yields

$$\mathbb{E}[I_i \mathbf{1}\{Y_i \in B\} \mid x] = \int \mathbf{1}\{y \in B\} \Pr(I_i = 1 \mid y, \ell, x) P_i(dy, d\ell \mid x). \tag{11}$$

Summing proves the identity. Normalization follows from $\sum_i I_i = k$ almost surely. No factorization of the joint law is used. □

When Q_x is absolutely continuous with respect to P_x^0, the mark-specific weight

$$r_x(y) = \frac{dQ_x}{dP_x^0}(y) \tag{12}$$

records overrepresentation or underrepresentation of semantic mark y. The main paper focuses on identity weights p_i/k because they are directly estimable from a complete panel trace and admit the exact decomposition below.

B.3 Exact identity and within-model decomposition

Theorem 2 (Identity-mixture and within-model decomposition)**.** *Let $\phi_i = \phi(Y_i)$ be integrable, $a_i = \mathbb{E}[\phi_i \mid x]$, $p_i = \mathbb{E}[I_i \mid x]$, and $c_i = \mathrm{Cov}(I_i, \phi_i \mid x)$. Then*

$$\Delta_\phi(x) = \mathbb{E}_{Q_x}[\phi(Y)] - \mathbb{E}_{P_x^0}[\phi(Y)] = \sum_i \left(\frac{p_i}{k} - \frac{1}{n}\right) a_i + \frac{1}{k}\sum_i c_i. \tag{13}$$

Proof. The selected-response expectation is

$$\mathbb{E}_{Q_x}[\phi(Y)] = \frac{1}{k}\sum_i \mathbb{E}[I_i \phi_i \mid x]. \tag{14}$$

By the definition of covariance,

$$\mathbb{E}[I_i \phi_i \mid x] = p_i a_i + c_i. \tag{15}$$

Subtracting $n^{-1}\sum_i a_i$ and collecting terms gives the result. □

The two components can be nonzero and cancel. Therefore, a zero total shift does not establish equal identity weights or independence within identities. Conversely, identity-independent random preselection sets $p_i = k/n$ and $c_i = 0$ for every i, so both components vanish separately at the selected-response level.

B.4 SHARP BOUNDS FOR UNRESOLVED ANSWERS

The selected-response shift remains partially identified when some model outputs do not resolve to a valid reference label. For one trace, let $Z_i \in \{0, 1\}$ denote the latent correctness of model i, let O be the identities with observed correctness, and let U be the unresolved identities. For a selected set S of size k, the item-level shift is the linear form

$$\Delta_Z = \sum_{i=1}^{n} \left(\frac{I_i}{k} - \frac{1}{n} \right) Z_i. \tag{16}$$

Each unresolved value can vary independently over $\{0, 1\}$. The sharp lower and upper bounds are therefore

$$\Delta_Z^{\mathrm{lo}} = \sum_{i \in O} \left(\frac{I_i}{k} - \frac{1}{n} \right) Z_i - \frac{|U \setminus S|}{n}, \tag{17}$$

$$\Delta_Z^{\mathrm{hi}} = \sum_{i \in O} \left(\frac{I_i}{k} - \frac{1}{n} \right) Z_i + \left(\frac{1}{k} - \frac{1}{n} \right) |U \cap S|. \tag{18}$$

The lower bound assigns every unresolved selected response to incorrect and every unresolved excluded response to correct. The upper bound uses the opposite assignment. These assignments attain the extrema because Equation 16 is separable in the unresolved values. Averaging either endpoint across traces preserves the bound, and the held-out analysis bootstraps the task-level averages of both endpoints.

B.5 VARIABLE SELECTED-SET SIZE

Some quorum policies select a random number of responses. Two sampling conventions then differ. Let $m_x = \mathbb{E}[|S| \mid x]$. Pooling all selected responses across runs gives the count-weighted law

$$Q_x^{\mathrm{count}}(B) = \frac{1}{m_x} \sum_i \int_{B \times \mathbb{R}_+} \pi_i(y, \ell \mid x) P_i(dy, d\ell \mid x). \tag{19}$$

Sampling one run first and then one response from that run gives

$$Q_x^{\mathrm{run}}(B) = \sum_i \int_{B \times \mathbb{R}_+} \lambda_i(y, \ell \mid x) P_i(dy, d\ell \mid x), \tag{20}$$

where

$$\lambda_i(y, \ell \mid x) = \mathbb{E}\left[\frac{I_i}{|S|} \middle| Y_i = y, L_i = \ell, X = x \right]. \tag{21}$$

Both reduce to Equation 3 for fixed k. The main selected-response analysis uses fixed first-three and therefore avoids this ambiguity.

B.6 EXACT NEUTRALITY CONDITION

Let P_i^Y denote the marginal mark law and define $\bar{\pi}_i(y \mid x) = \Pr(I_i = 1 \mid Y_i = y, X = x)$. Integrating Equation 3 over completion time shows that $Q_x = P_x^0$ exactly when

$$\sum_i \bar{\pi}_i(y \mid x) P_i^Y(dy \mid x) = \frac{k}{n} \sum_i P_i^Y(dy \mid x) \tag{22}$$

as measures. The condition $\bar{\pi}_i(y \mid x) = k/n$ for every identity and mark is sufficient. It is stronger than necessary because different identity-specific deviations can cancel in the mixture.

Within-model independence $Y_i \perp L_i \mid X$ does not imply Equation 22 for a heterogeneous panel. Different latency distributions can create unequal p_i, and different identities can have different P_i^Y. This is the reason an identity term remains even when every within-model covariance is zero.

B.7 JOINT DECISION LAW

Theorem 3 (Joint decision law)**.** *Let $\Pi_{\Gamma,x}(s \mid y, \ell)$ be the policy's conditional selected-subset kernel. For any measurable decision set C,*

$$\Pr_{\Gamma}(D \in C \mid x) = \sum_{s} \int \mathbf{1}\{A(y_s) \in C\} \Pi_{\Gamma,x}(s \mid y, \ell) P_x(dy, d\ell). \tag{23}$$

Proof. Condition on the complete potential trace $(Y_{1:n}, L_{1:n}) = (y, \ell)$ and then on $S = s$. The decision is the deterministic value $A(y_s)$ once these variables are fixed. Integrating the conditional indicator over the subset kernel and the joint trace law gives the identity. □

The selected-response marginal cannot replace the joint law. Consider three binary marks with $\Pr(Y_i = 1) = 0.6$ for every identity. Under independence, majority is one with probability $0.6^3 + 3(0.6)^2(0.4) = 0.648$. Under perfect correlation, majority is one with probability 0.6. The one-item marginals agree and the majority laws differ.

For two policies applied to the same potential trace, paired disagreement is

$$d_{\Gamma,\Gamma_0}(x) = \Pr(D_\Gamma \neq D_{\Gamma_0} \mid X = x). \tag{24}$$

This common-trace construction is one coupling of the two decision laws. The coupling inequality gives the following total variation (TV) bound:

$$\|\mathcal{L}(D_\Gamma \mid x) - \mathcal{L}(D_{\Gamma_0} \mid x)\|_{\mathrm{TV}} \leq d_{\Gamma,\Gamma_0}(x). \tag{25}$$

The empirical first-three/full disagreement rate is therefore a pathwise policy comparison under the observed coupling. It upper-bounds decision-law total variation, but equality need not hold.

B.8 STATIC FIRST-K INCLUSION PROBABILITY

The general propensity law does not require independence. A closed form is available under independent, nonidentical response-time pairs and continuous completion times. For focal identity i with $L_i = \ell$, define $F_j(\ell \mid x) = \Pr(L_j < \ell \mid x)$. Model i enters first-k exactly when at most $k - 1$ other models finish earlier. Hence

$$w_i(\ell \mid x) = \Pr\left(\sum_{j \neq i} \mathbf{1}\{L_j < \ell\} \leq k - 1 \middle| x \right), \tag{26}$$

which is the Poisson-binomial cumulative probability

$$w_i(\ell \mid x) = \sum_{A \subseteq [n] \setminus \{i\},\, |A| \leq k-1} \prod_{j \in A} F_j(\ell \mid x) \prod_{j \notin A,\, j \neq i} (1 - F_j(\ell \mid x)). \tag{27}$$

Substituting w_i for π_i in Equation 3 gives a computable selected-response law.

In the identically distributed case,

$$w_{n,k}(\ell) = \sum_{r=0}^{k-1} \binom{n-1}{r} F_L(\ell)^r (1 - F_L(\ell))^{n-1-r}, \tag{28}$$

and rank symmetry gives $\mathbb{E}[w_{n,k}(L)] = k/n$. The feature shift becomes

$$\Delta_\phi = \frac{n}{k} \operatorname{Cov}(\phi(Y), w_{n,k}(L)). \tag{29}$$

This special case explains how within-model response-time dependence can set the sign of a shift. It does not describe the heterogeneous identity effect observed in the main experiment.

B.9 TARGET-SET DELAY BUDGET

Theorem 4 (Delay budget infimum)**.** *For a realized finite trace, static first-k, and a fixed mark-independent tie rule, the infimum total nonnegative delay required to make target subset S arrive first is*

$$B_{\inf}(S) = \sum_{i \notin S} [t_S - L_i]_+, \qquad t_S = \max_{j \in S} L_j. \tag{30}$$

Proof. Every target member finishes no earlier than its original time, so the threshold at which all target members are visible is at least t_S. Each outsider with $L_i < t_S$ must be delayed to at least this threshold, which requires $\delta_i \geq t_S - L_i$. Outsiders already at or above the threshold require no delay. Summing gives the lower bound.

For achievability of the infimum, leave every target member unchanged. For any $\epsilon > 0$, delay each outsider with $L_i \leq t_S$ to $t_S + \epsilon$ and leave later outsiders unchanged. The target set is then strictly first, and the total added delay tends to the displayed lower bound as $\epsilon \downarrow 0$. A favorable fixed tie order can attain the bound exactly. Otherwise, the value is an infimum. □

Let $D_0 = A(Y_{S_k(L)})$ be the natural first-k decision. The budget infimum for target decision a is

$$B^*(a) = \min_{S:|S|=k,\, A(Y_S)=a} B_{\inf}(S), \tag{31}$$

and the decision-flip margin is $M_{\text{flip}} = \min_{a \neq D_0} B^*(a)$. If all completion times are finite and the intervention permits arbitrary finite nonnegative delay, every fixed subset has finite $B_{\inf}$. Finite decision-flip reachability is then equivalent to the existence of at least one subset with another aggregate decision. Completion times determine the margin magnitude, while the realized marks and aggregator determine whether another decision exists.

B.10 COMPLETION-INVARIANT COMMIT

Theorem 5 (Sufficiency and necessity of completion invariance)**.** *Suppose the realized unseen continuation belongs to $\mathfrak{C}(h_t)$. If $|\mathcal{D}(h_t)| = 1$, committing to its unique element agrees with the full-completion decision. If $|\mathcal{D}(h_t)| > 1$, no deterministic rule using only h_t can guarantee agreement for every admissible completion.*

Proof. For sufficiency, the realized continuation c^* is one member of $\mathfrak{C}(h_t)$, so $A(h_t \oplus c^*)$ equals the unique element of $\mathcal{D}(h_t)$. For necessity, choose $c_1, c_2 \in \mathfrak{C}(h_t)$ with different aggregate decisions. Both worlds have the same visible history at time t. A deterministic current rule emits one value in both worlds and must disagree with at least one full decision. □

For binary majority with odd total size n, let the visible counts be a and b and let $r = n - a - b$ marks remain unseen. Every completion yields the same majority exactly when $|a - b| > r$. The multiclass implementation used in the experiment enumerates equality patterns among unseen labels and applies the singleton-decision definition directly. The binary condition is only a special case.

Proposition 6 (Extremal plurality certificate)**.** *Assume a nonempty visible prefix and that every assignment from a common mark alphabet is admissible. Plurality ties are resolved by the highest-priority supporting identity under an order fixed before any mark is observed. Let E_t contain the m visible labels and let R_t contain the r unseen identities. For each $a \in E_t$, let $c^{(a)}$ assign every identity in R_t to label a. When a label outside E_t is admissible, also let $c^{(\star)}$ assign them all to one such label. Then $|\mathcal{D}(h_t)| = 1$ if and only if*

$$A(h_t \oplus c) = w \qquad \textit{for every such extremal completion } c \tag{32}$$

for the same visible label w.

Proof. Assume label a wins under an arbitrary continuation. Reassigning every unseen identity to a weakly increases its count and can only improve the priority of its best supporting identity. Every competing label weakly loses votes and can only lose supporting priority. Hence a also wins under the extremal continuation $c^{(a)}$. The same argument applies to a novel winning label through $c^{(\star)}$.

Therefore, if every extremal continuation returns the same visible label w, no other continuation can return another label. The converse follows because each extremal continuation is admissible. □

The exact finite implementation enumerates only admissible label equalities. For plurality over a fixed finite panel, every unseen equality pattern can be represented using the labels already visible plus one fresh label for each unseen identity. The direct implementation evaluates at most $(m + r)^r$ assignments at a prefix. Proposition 6 gives an equivalent implementation with at most $m + 1$ extremal completions. Cached counts, the best supporting priority for each label, and the two leading competitors reduce the per-prefix test to $O(n + m)$ time and $O(m)$ memory. The experiment uses direct enumeration at $n = 5$. An independent implementation audit compares both procedures on 5,286 canonical prefix states through $n = 7$ and 10,000 additional states with panels of size 8 through 16, with zero disagreements.

The delay target calculation is also finite. It enumerates all $\binom{n}{k}$ target subsets, computes t_S and $B_{\inf}(S)$, applies the aggregator to each subset, and chooses the admissible target specified by the intervention objective. Its time complexity is $O(n\binom{n}{k})$ after the trace is available. These two procedures implement Equations 6 and 7 directly.

B.11 Selected-only non-identifiability

Selected-only logs generally cannot recover the nominal all-model mixture. Consider $n = 2$ and $k = 1$. In world A, both marks equal one. In world B, the faster mark equals one and the slower mark equals zero. The first-response mark and time can have the same law in both worlds, while P_x^0 differs. Therefore, no estimator based only on the selected trace identifies P_x^0 over arbitrary joint laws. Shadow completion, pre-generation identity randomization, known positive propensities, or an additional structural model supplies the missing information.

C Experimental protocol

C.1 Model panel

Table 3 lists the panel, which spans 0.5B to 8B parameters. All checkpoints are instruction-tuned public releases and remain frozen during evaluation.

Table 3: Model panel and overall measured behavior. Accuracy is the base answer accuracy before selection. Inclusion is the fraction of first-three selected sets containing the identity.

Model	Parameters	Accuracy (%)	First-three inclusion (%)
Qwen2.5 Instruct	0.5B	22.92	85.07
Qwen2.5 Instruct	1.5B	52.08	86.11
Qwen3	0.6B	50.00	65.28
Gemma 2 Instruct	2B	47.92	63.54
Llama 3.1 Instruct	8B	59.72	0.00

C.2 Task selection

We use the public test splits of GSM8K and SVAMP and the public MATH-500 release introduced by Lightman et al. (2023). MATH examples are eligible when the reference answer is a signed integer, decimal, or reduced fraction that can be compared exactly. Seed 20260821 selects 48 examples without replacement from each eligible pool. Selection occurs before model generation. Table 4 summarizes the resulting 144 inference clusters and comparison rules.

C.3 Prompt, decoding, and marks

Each checkpoint receives one user-role message formed from three blocks. No separate system-role message is used. The exact fixed text and task insertion are:

Table 4: Benchmark sampling and evaluation.

Benchmark	Public pool	Selected	Answer comparison
GSM8K	1,319	48	exact normalized number
SVAMP	300	48	exact normalized number
MATH-500 numeric	366 eligible	48	exact integer, decimal, or fraction

```
You solve exact quantitative problems.  End with
exactly one line in the form FINAL: <number>.  The
number may be an integer, decimal, or reduced
fraction.

Solve the problem carefully.  Do not omit the final
answer.

Question:  <benchmark question>
```

Decoding is greedy with temperature zero, top-p one, and a 256-token completion limit. Each model-task pair is executed twice. The second pass is a technical repeat, not an independent semantic sample.

The parser first checks the requested final line and then applies benchmark-compatible numeric normalization. A response that stops at a valid final-answer endpoint receives its normalized numeric label. Any invalid endpoint maps to the explicit mark `ABSTAIN`. A number embedded in unfinished reasoning never becomes a protocol vote. Correctness is defined by equality with the normalized reference answer. The primary protocol keeps every sampled request, including abstention marks. All-valid-endpoint and all-end-of-sequence subsets are sensitivity analyses.

C.4 Timing and shadow construction

Models are evaluated separately over the same fixed task order. Per-sequence completion service time is recorded by accelerator events within each model's batch. A virtual panel starts all five measured sequences at time zero and orders their terminal marks by service time. This construction estimates a model-isolated worker regime. It preserves every potential mark, so full-five, arrival first-three, every preselected subset, and the decision-equivalent policies are deterministic functions of the same trace.

The shadow design does not estimate queueing or resource feedback among colocated models, and it excludes cancellation because all responses are retained. Holding outputs and service times fixed isolates changes caused by policy replay.

C.5 Aggregation and tie handling

Plurality counts normalized protocol marks. A fixed identity priority resolves equal vote counts. The priority order, sampled from seed 20260821 before any output or latency is observed, is Qwen2.5 0.5B, Gemma 2 2B, Llama 3.1 8B, Qwen3 0.6B, and Qwen2.5 1.5B. It remains constant across all policies. Arrival first-three selects the three smallest measured completion times, with the same priority resolving exact latency ties. Full-five uses all marks.

Uniform random-three is evaluated without Monte Carlo subset sampling. We enumerate all $\binom{5}{3} = 10$ identity subsets and average their selected-response and decision outcomes with weight $1/10$. This calculation sets each identity's inclusion probability to $3/5$ exactly.

Completion-invariant plurality processes marks in arrival order. After each arrival, it enumerates all equality patterns for the unseen labels, applies full-five plurality under the fixed tie rule, and commits when all admissible patterns yield one decision. Quorum-3 else full commits as soon as one label has three observed votes and otherwise uses full-five plurality.

C.6 Estimands and uncertainty

For trace r, let Z_r^{item} be first-three mean item correctness minus full-five mean item correctness. Let Z_r^{decision} be first-three decision correctness minus full-five decision correctness. Relative latency reduction is

$$Z_r^{\text{latency}} = 1 - \frac{L_{r,\text{commit}}}{\max_i L_{ri}}. \tag{33}$$

The two technical repeats are averaged within a task. We then average over the 144 task clusters. Confidence intervals resample tasks with replacement 10,000 times from the pooled task set; benchmark-specific intervals resample the corresponding tasks.

For the decomposition, ϕ_i is correctness. The empirical a_i is each identity's mean correctness, p_i is its first-three inclusion rate, and c_i is the empirical covariance between inclusion and correctness. Each bootstrap replicate recomputes all three terms after task resampling and checks their exact equality.

C.7 Sample size and achieved precision

The sample size is fixed at 48 tasks per benchmark, giving 144 task clusters. All sampled tasks remain in the analysis. The observed task-cluster standard deviation of the selected-response contrast is 0.1452, corresponding to a normal-approximation half-width of $1.96 \times 0.1452/\sqrt{144} = 0.0237$. This achieved-precision diagnostic complements the reported bootstrap intervals; it is calculated after evaluation and does not determine the sample size.

D Complete experimental results

D.1 Model-level decomposition

Table 5 shows each identity's contribution to Equation 4, and Table 6 reports the corresponding overall estimate and task-cluster intervals. The Llama identity has the largest negative identity contribution because it combines the highest base accuracy with zero first-three inclusion. Its within-model contribution is zero because inclusion is constant.

Table 5: Per-identity terms for the selected-response accuracy decomposition. Contributions are proportions. Multiplying by 100 gives percentage points.

Model	Inclusion	Accuracy	Identity term	Within-model term
Qwen2.5 0.5B	0.8507	0.2292	0.01915	−0.01174
Qwen2.5 1.5B	0.8611	0.5208	0.04533	−0.01292
Qwen3 0.6B	0.6528	0.5000	0.00880	0.00231
Gemma 2 2B	0.6354	0.4792	0.00566	0.01194
Llama 3.1 8B	0.0000	0.5972	−0.11944	0.00000
Sum	3.0000		−0.04051	−0.01042

Table 6: Exact decomposition with task-cluster bootstrap intervals.

Term	Estimate	95% low	95% high
Observed selected-response shift	−0.05093	−0.07455	−0.02731
Identity-mixture term	−0.04051	−0.05456	−0.02699
Within-model covariance term	−0.01042	−0.02856	0.00786

The terms sum to -0.05093, with an absolute residual below 5×10^{-17}. This finite-sample equality is an arithmetic check of the theorem and uses no fitted approximation.

D.2 Policy decomposition and decision-equivalent policies

Table 7 reports the paired policy contrasts and decision-equivalent latency reductions used in the main comparison. Arrival first-three minus full-five equals the sum of arrival first-three minus random-three and random-three minus full-five for every task-level contrast before averaging. This exact equality follows from using the same complete trace and exhaustive subset baseline.

Table 7: Paired overall effects used in the main policy comparison. The last two columns give 95% task-cluster bootstrap interval endpoints.

Contrast or metric	Estimate	95% low	95% high
Arrival-three minus random-three decision accuracy	−0.08125	−0.12708	−0.03681
Random-three minus full-five decision accuracy	−0.06111	−0.09514	−0.03056
Arrival-three minus full-five decision accuracy	−0.14236	−0.20486	−0.08333
Random-three minus full-five item accuracy	0.00000	0.00000	0.00000
Arrival-three latency reduction	0.67139	0.65364	0.68902
Random-three expected latency reduction	0.24164	0.23323	0.24998
Completion-invariant latency reduction	0.40829	0.35927	0.45771
Quorum-3 else full latency reduction	0.35567	0.30334	0.40764

Table 8: Commit depth on the 288 complete traces. Each entry counts traces by the number of observed responses at commit.

Policy	Three responses	Four responses	Five responses
Full-five plurality	0	0	288
Arrival first-three	288	0	0
Quorum-3 else full	88	78	122
Completion-invariant	88	104	96

D.3 Finite policy and tie-rule sensitivity

Table 9 evaluates every cardinality in the static first-k family on the same complete traces. The selected-response shift decreases in magnitude as k approaches the full panel, while the latency reduction also decreases. Decision accuracy need not move monotonically with k because plurality and its tie rule are nonlinear.

Table 9: Exact first-k sensitivity on the 288 complete traces. All entries except k are percentages or percentage points.

k	Response shift	Decision shift	Disagreement	Latency reduction
1	−14.58	−25.00	46.18	85.17
2	−7.12	−28.12	37.85	75.13
3	−5.09	−14.24	22.92	67.14
4	−2.95	−4.86	9.03	58.17
5	0.00	0.00	0.00	0.00

The aggregator fixes one identity priority before observing any output. To test dependence on this deterministic choice, Table 10 exhausts all $5! = 120$ fixed priority orders. Every order retains a negative arrival-minus-full shift and a negative arrival-minus-random-three shift. Thus, the sign of the first-three result does not depend on the chosen order, although its magnitude and the disagreement rate vary.

The selected-response shift is negative for every $k < 5$ and reaches zero at the full panel. At $k = 3$, both decision contrasts remain negative under all 120 tie orders. Panel composition, runtime mapping, and task domain remain fixed in these comparisons.

Table 10: Exhaustive sensitivity over all 120 fixed identity tie-priority orders. The random-three value averages all ten identity subsets under each priority order.

Quantity	Minimum	Median	Maximum	Primary tie order
Arrival minus full accuracy (pp)	−20.14	−11.11	−7.99	−14.24
Arrival minus random-three (pp)	−13.40	−6.35	−1.74	−8.12
First-three/full disagreement (%)	17.36	23.09	29.86	22.92

D.4 Benchmark-level results

Table 11 reports the complete benchmark-level estimates and intervals. The MATH numeric subset has less decision movement despite similar latency reduction because its observed marks more often leave the plurality decision unchanged under membership changes. This contrast supports the boundary analysis and shows why latency reduction alone does not determine decision quality.

Table 11: Primary protocol effects by benchmark. Each cell reports a proportion above its 95% task-cluster bootstrap CI in parentheses.

Benchmark	Selected-response shift	Decision shift	Latency reduction
GSM8K	−0.0736 (−0.1153, −0.0319)	−0.1458 (−0.2500, −0.0625)	0.6558 (0.6296, 0.6827)
SVAMP	−0.0708 (−0.1181, −0.0222)	−0.2604 (−0.3958, −0.1250)	0.7363 (0.7054, 0.7664)
MATH numeric	−0.0083 (−0.0361, 0.0208)	−0.0208 (−0.1042, 0.0417)	0.6221 (0.5975, 0.6486)
All	−0.0509 (−0.0745, −0.0273)	−0.1424 (−0.2049, −0.0833)	0.6714 (0.6536, 0.6890)

D.5 Boundary localization

Table 12 separates the overlapping boundary, diversity, and endpoint diagnostics. The four tie-free disagreements have equal decision correctness under both policies even though the answer label changes. Their contribution to decision accuracy shift is therefore zero. The 62 boundary disagreements contain the observed correctness changes that drive the aggregate decision effect.

Table 12: Natural first-three versus full-five effects by diagnostic scope. Scopes overlap and should not be summed.

Scope	Traces	Item shift	Decision shift	Disagreements
Tie or plurality boundary	96	−0.1528	−0.4271	62
Tie-free	192	0.0000	0.0000	4
Semantically diverse	224	−0.0655	−0.1830	66
Semantically uniform	64	0.0000	0.0000	0
All endpoints valid	122	−0.1115	−0.2541	36
All end-of-sequence	122	−0.1115	−0.2541	36

D.6 Endpoint sensitivity

Table 13 compares the primary protocol with endpoint-conditioned sensitivities. The strict complete-case result is conditional on endpoint eligibility and is not used as an estimate for all sampled requests. Its role is to test whether invalid endpoints explain the primary protocol effect. The effect remains and increases in magnitude within this eligible subset.

Table 13: All-valid-endpoint sensitivity. The task-cluster count is lower because complete-case eligibility depends on all five responses.

Scope	Traces	Task clusters	Item shift	Decision shift
Primary, all sampled	288	144	-0.0509	-0.1424
All answer endpoints valid	122	61	-0.1115	-0.2541
All end-of-sequence	122	61	-0.1115	-0.2541
First-valid-three vs. full-valid	288	144	-0.0576	-0.1285

D.7 Execution stability

Greedy decoding produces identical text and normalized answers across both executions for all 720 model-task pairs. Across the resulting 144 task panels, 137 preserve the exact five-model arrival rank, giving 95.14% rank stability. The task-cluster analysis averages executions before resampling, so technical repetition does not inflate the inferential sample size. Table 14 reports execution stability and data completeness.

Table 14: Execution stability and data-completeness checks.

Check	Result
Expected model generations	1,440 / 1,440
Expected five-model traces	288 / 288
Exact greedy text agreement	100.00%
Normalized answer agreement	100.00%
Exact complete arrival-rank agreement	95.14%
Completion-invariant/full mismatches	0 / 288
Quorum-3/full mismatches	0 / 288
Verified constructive-delay failures	0 / 128

D.8 Deterministic disagreement cases

Table 15 gives the lexicographically first trace in every observed benchmark by outcome-class cell. This deterministic rule makes the case selection reproducible.

Table 15: Deterministically selected natural disagreements. `ABSTAIN` is a protocol mark and can win a plurality under the fixed tie rule.

Benchmark	Outcome class	Ground truth	First-three	Full-five
GSM8K	both wrong, different	112	320	`ABSTAIN`
GSM8K	full correct	163	80	163
MATH	both wrong, different	$1/3$	$1/2$	`ABSTAIN`
MATH	full correct	9901	1980	9901
MATH	first-three correct	56	56	`ABSTAIN`
SVAMP	both wrong, different	527292	539952	520002
SVAMP	full correct	1	-6	1
SVAMP	first-three correct	105	105	`ABSTAIN`

D.9 Direction of reachable changes

Table 16 reports the specified target directions and their constructive budgets. Reachability is an intervention property: with unbounded finite delay, any target subset can be made first by delaying outsiders. The budget distribution and controlled replay measure how much delay was required in this panel.

Table 16: Ingress target directions and constructive budgets. Fractions condition on the natural first-three correctness state.

Direction	Reachable	Conditional fraction	Median budget (ms)
Correct to wrong	23	23 / 123	1,351
Wrong to correct	65	65 / 165	4,461
Wrong to another wrong	40	40 / 165	5,865
Any decision change	128	128 / 288	4,476

E HELD-OUT MULTIPLE-CHOICE CONFIRMATION

E.1 MODELS, BENCHMARKS, AND FROZEN PROTOCOL

The held-out study changes the full model panel and moves from numeric reasoning to four multiple-choice benchmarks. The panel contains OLMo 2 1B, SmolLM2 1.7B, Granite 3.3 2B, Phi-3.5 Mini, and Mistral 7B. Table 17 reports the measured identity statistics. The two fastest identities enter every first-three set and have the two lowest base accuracies. Phi-3.5 Mini has the highest base accuracy and enters 2.08% of first-three sets. Figure 5 summarizes the benchmark-level bounds, runtime mappings, and decomposition developed below.

Table 17: Held-out panel behavior under native measured service time. Accuracy is computed before arrival selection. Mean service time is measured per completed sequence.

Model	Size	Base accuracy (%)	First-three inclusion (%)	Mean time (s)
OLMo 2 Instruct	1B	23.96	100.00	2.46
SmolLM2 Instruct	1.7B	43.75	100.00	3.52
Granite 3.3 Instruct	2B	61.98	84.38	9.66
Phi-3.5 Mini Instruct	3.8B	79.17	2.08	21.19
Mistral 7B Instruct	7B	70.83	13.54	16.53

The task sample contains 48 examples from each of ARC-Challenge test, the full MMLU test mixture, TruthfulQA multiple-choice validation, and HellaSwag validation. The 192 sampled task identities are disjoint from the sample used to develop the answer parser. Two greedy executions of each model-task pair yield 384 complete five-model traces and 1,920 generations. The same fixed identity priority resolves plurality ties in every policy and runtime mapping. We call the measured, unmodified service times native time.

The primary parser was frozen before held-out generation, without using answer correctness. It first returns the last valid case-insensitive occurrence of `FINAL: <LETTER>`. If no such occurrence exists, it accepts a valid option label at the start of the first nonempty line when followed by a period, parenthesis, colon, or comma. Every other output receives the operational abstention mark. This parser resolves 96.35% of generation rows. The strict sensitivity parser accepts only an exact final nonempty line of the form `FINAL: <LETTER>`.

The inference unit is the task identity. Each statistic is averaged over the two executions within a task. We then obtain 95% confidence intervals (CIs) from 10,000 percentile bootstrap resamples of the 192 task identities. The observed protocol retains abstentions as semantic marks. Equations 16 and 18 also evaluate every unresolved answer assignment without assumptions about its correctness.

E.2 BENCHMARK-LEVEL CONFIRMATION

Arrival first-three lowers selected-response accuracy on all four held-out benchmarks. Table 18 reports the observed shift and its sharp upper bound under unresolved-answer assignments. Every observed CI and every upper-bound CI remains below zero. Pooled across benchmarks, the observed loss is 11.84 points (95% CI, 9.86 to 13.82). Even the assignment most favorable to arrival first-three leaves a 10.17-point loss (95% CI, 8.02 to 12.29).

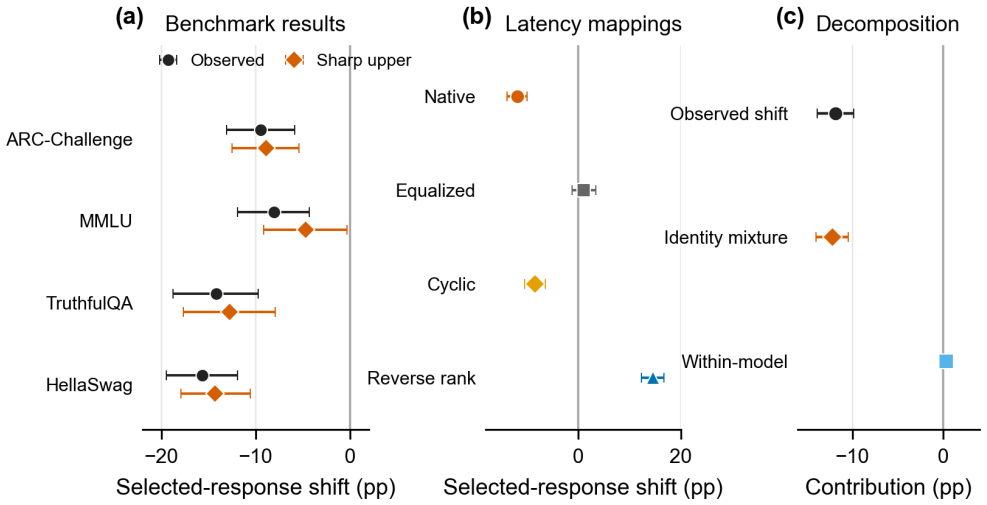

Figure 5: The held-out panel reproduces identity-driven arrival selection. (a) Observed selected-response shifts and sharp unresolved-answer upper bounds remain negative on all four benchmarks. (b) Response-preserving runtime mappings move or reverse the shift by changing which identities arrive first. (c) Under native time, the identity-mixture term accounts for the observed shift, while the within-model interval includes zero. Lines show 95% task-cluster bootstrap intervals.

Table 18: Held-out effects under native service time. Each cell reports a percentage-point estimate above its 95% task-cluster bootstrap CI in parentheses. The sharp upper column maximizes the selected-response shift over all unresolved-answer assignments.

Benchmark	Observed selected shift	Sharp upper shift	Decision shift
ARC-Challenge	−9.44 (−13.06, −5.83)	−8.89 (−12.50, −5.42)	−18.75 (−31.25, −8.33)
MMLU	−8.06 (−11.94, −4.31)	−4.72 (−9.17, −0.28)	−10.42 (−20.83, 0.00)
TruthfulQA	−14.17 (−18.75, −9.72)	−12.78 (−17.64, −7.92)	−27.08 (−41.67, −14.58)
HellaSwag	−15.69 (−19.44, −11.94)	−14.31 (−17.92, −10.56)	−33.33 (−47.92, −18.75)
All	−11.84 (−13.82, −9.86)	−10.17 (−12.29, −8.02)	−22.40 (−28.65, −16.15)

The decision comparison reaches the same conclusion at the nonlinear aggregator. Arrival first-three disagrees with full-five plurality on 29.69% of traces and lowers decision accuracy by 22.40 points. Exhaustively averaging all ten response-independent three-model subsets yields 62.03% decision accuracy, compared with 46.88% for arrival first-three. Their 15.16-point difference (95% CI, 10.57 to 19.79) isolates arrival-conditioned subset selection from the use of three votes.

E.3 Policies and decision equivalence

Completion-invariant commit gives the larger latency reduction among the two decision-equivalent policies in the held-out panel. It agrees with all 384 full-five decisions and reduces mean latency by 22.49% (95% CI, 19.27 to 25.85). Quorum-3 else full is also exact and reduces mean latency by 18.15%. Table 19 reports every evaluated policy under native service time.

Arrival stability commits at the first prefix whose fixed-priority plurality decision repeats and has at least two supporting marks; otherwise it waits for all five. The two-support requirement is fixed

for every task. This policy provides a latency-accuracy operating point, while completion invariance provides an exact full-decision guarantee.

Table 19: Held-out policy results under native service time. Disagreement compares each policy with full-five plurality.

Policy	Decision accuracy (%)	Disagreement (%)	Mean selected	Latency reduction (%)
Full five	69.27	0.00	5.00	0.00
Uniform random-three	62.03	21.35	3.00	15.19
Arrival first-three	46.88	29.69	3.00	63.37
Arrival stability	63.54	6.25	3.51	43.96
Quorum-3 else full	69.27	0.00	4.41	18.15
Completion-invariant	69.27	0.00	4.28	**22.49**

E.4 Runtime mappings and exact decomposition

Runtime remapping tests the functional dependence in Equation 3 while preserving every response. Identity-median equalization rescales each model's trace times so all identity medians coincide. The fixed cyclic mapping assigns each identity another identity's within-trace times using a fixed cycle. The reverse-rank mapping assigns the fastest identity's time trace to the slowest identity, the second fastest to the second slowest, and so on. Both permutation mappings preserve the complete multiset of five service times within every trace.

Table 20 shows that the selected-response shift follows the induced identity mixture. Equalizing identity medians changes the native 11.84-point loss to a 1.09-point gain whose interval includes zero. Fixed cyclic reassignment produces an 8.37-point loss, whereas reversing median rank produces a 14.55-point gain. In every mapping, Equation 4 reconstructs the observed shift to numerical precision, and the identity-mixture term determines the large sign changes.

Table 20: Exact held-out decomposition under four response-preserving runtime mappings. Each cell reports a percentage-point estimate above its 95% task-cluster bootstrap CI in parentheses.

Runtime mapping	Selected shift	Identity mixture	Within-model term
Native	-11.84 $(-13.82, -9.86)$	-12.19 $(-13.98, -10.41)$	0.35 $(-0.30, 1.10)$
Median equalized	1.09 $(-1.16, 3.42)$	-0.27 $(-1.53, 1.01)$	1.37 $(-0.58, 3.31)$
Cyclic	-8.37 $(-10.42, -6.35)$	-7.00 $(-8.98, -5.05)$	-1.37 $(-2.36, -0.43)$
Reverse rank	14.55 $(12.29, 16.74)$	13.64 $(11.60, 15.72)$	0.91 $(-0.05, 1.89)$

E.5 Parser sensitivity

The parser sensitivity separates a usable endpoint rule from a low-coverage grammar. Table 21 reports both prespecified parsers. The frozen primary rule resolves 96.35% of outputs, and its complete unresolved-answer interval remains negative. The strict rule resolves only 23.54% of outputs. Its sharp interval spans 62.57 points and therefore does not identify the direction of the selected-response effect.

Because the strict parser leaves more than three quarters of latent answers unresolved, its sharp interval spans both signs and does not identify a semantic direction. Under the primary parser, the result remains negative under every assignment of its 3.65% unresolved rows: even the assignment that maximizes the shift leaves the pooled and all four benchmark intervals below zero.

Table 21: Held-out parser sensitivity under native service time. Observed shifts include abstention as an operational mark. The sharp lower and upper columns vary every unresolved answer over all valid options.

Parser	Endpoint rate (%)	Observed shift (pp)	Sharp lower (pp)	Sharp upper (pp)	Decision shift (pp)
Frozen primary	96.35	−11.84	−12.99	−10.17	−22.40
Strict final line	23.54	3.65	−31.15	31.42	1.56

F Payload-fixed ingress replay

F.1 Controller construction

The ingress experiment begins from the complete set of 288 shadow traces. Each trace supplies five response payloads, five baseline delivery times, a fixed identity tie order, and the first-three aggregator. An ordered TCP replay sends all five identity-tagged payloads over one connection to the receiver. Baseline replay uses measured virtual delivery times. Intervention replay adds the constructive nonnegative delays derived from Equation 6 and sorts the resulting schedule before transmission.

The sender waits between messages using a common positive scale factor between virtual milliseconds and wall seconds. TCP preserves the scheduled message order; host scheduling can change elapsed wall time. The reported budget uses virtual delay, not measured socket latency. Each identity's serialized payload is sent once, and the receiver records payload identity, arrival order, selected set, and final decision.

F.2 Target construction

For every trace, all ten three-identity subsets are enumerated. The baseline decision uses the measured first-three subset. When that decision is wrong and at least one subset returns the ground truth, the target is the lowest-budget correct subset. Otherwise, the target is the lowest-budget subset with any different decision. This target rule prioritizes corrective changes whenever they are reachable. On 18 traces, it selects a higher-budget correct target instead of the globally closest alternative decision. If an outsider lies exactly at the target threshold and the fixed tie order is unfavorable, the constructive schedule uses the smallest representable later time. The target subset is then recomputed from the scheduled times before network replay.

This process identifies 128 traces with a reachable alternative decision. The other 160 traces have the same aggregate decision for all ten subsets and therefore have infinite decision-flip margin under the definition, even though every subset remains temporally reachable.

F.3 Verification results

The baseline stage verifies all 288 received orders, selected subsets, and decisions against the source traces. The intervention stage verifies all 128 target orders, subsets, and decisions. For every intervention, the serialized response associated with each identity is byte-identical to its baseline payload. Table 22 summarizes these checks.

The replay tests the static first-three controller under constructed post-generation schedules. It realizes the predicted semantic changes with identical payloads and verifies their ingestion through TCP. The transmission order is controlled by the sender, rather than emerging from independently delayed network clients. Generation feedback, shared accelerator scheduling, cancellation, and adaptive public-network attacks are outside this replay.

Table 22: Ingress replay verification.

Verification item	Passed
Baseline trace count	288 / 288
Baseline arrival order and decision	288 / 288
Mathematical target construction	128 / 128
Intervention received order	128 / 128
Intervention selected subset	128 / 128
Intervention target decision	128 / 128
Payload identity for every model	128 / 128

G Paired execution on a shared GPU

G.1 Models, workloads, and paired timing

The paired study measures how shared execution changes arrival-conditioned voting. The five reasoning models remain resident on one 80 GB GPU throughout both conditions. We reuse the 144 reasoning tasks and the fixed prompt, numeric parser, and identity priority from Appendix C. Each workload is a batch of 24 tasks. Six workloads partition the task set, and each contains eight tasks from each benchmark. Within each benchmark, four tasks participated in the deployment pilot and four did not. Seed 20260908 fixes the partition using task identities before the new execution. Membership does not depend on response quality or timing. The 72 nonpilot tasks were used in the original reasoning study, so this split tests transfer beyond the deployment pilot rather than a new benchmark sample.

Both conditions use the same generation and memory settings. Native model generation uses bfloat16 arithmetic, scaled dot-product attention, dynamic key-value caching, one beam, and greedy token selection with a 256-token completion budget. Explicit call arguments disable sampling and model-default overrides. Qwen3 uses its non-thinking chat template. The effective generation configuration is retained for each model. All sampled tasks remain in the analysis, with invalid endpoints represented by the existing abstention mark.

Shared execution starts five model workers from a common synchronization barrier. Each worker uses a separate accelerator stream, and a host-clock callback records each response at its first end-of-sequence token or token-budget completion. Isolated execution runs the same workers one at a time with the same model residency, task batch, and decoder. Each worker has its own start time. Tokenization and model loading precede timing. Both conditions use batch size 24 without memory-driven splitting. Two technical repetitions alternate condition order within each workload, yielding 2,880 generations and 288 complete panel traces per condition. The isolated condition defines virtual simultaneous starts from the individual worker measurements. Shared execution includes actual resource contention and host scheduling.

The policy comparison uses observed response completion times. We apply arrival first-three, arrival stability, quorum-3 else full, completion-invariant commit, and full-five plurality to each complete trace. Arrival stability uses the rule in Appendix E. We retain all completions and compute submission time from the last selected response. Relative commit reduction therefore measures an earlier decision under full collection, with no cancellation or computational savings assumed.

G.2 Workload-level uncertainty

Uncertainty is computed at the workload level because tasks in a batch share execution conditions. Each statistic first yields six workload means, with both repetitions of every task kept together. We resample these means 4,000 times to form percentile intervals. Benchmark and pilot-membership summaries use their corresponding tasks within each workload, with equal counts in every workload. Table 25 also reports each workload directly. These six clusters describe variability within the fixed task pool on one device. Generalization across production workloads and machines requires broader sampling.

G.3 DECISION-PRESERVING COMMIT UNDER CONTENTION

Completion invariance preserves every full decision: all 576 panel traces agree across the two conditions. Each model also produces identical output tokens across conditions and technical repetitions. In shared execution, it reduces mean commit time by 12.91%, with a 95% workload-cluster interval from 10.90% to 14.93%. Quorum-3 else full also preserves every full decision and reduces commit time by 11.59%. Table 23 reports all policies under both conditions.

Table 23: Policies under paired execution on 144 tasks. Each condition has 288 traces in six workloads. Disagreement is relative to full five. Commit reduction is the mean percentage reduction in replayed submission time, followed by its 95% workload-cluster bootstrap interval in parentheses.

Policy	Accuracy (%)	Disagreement (%)	Commit reduction (%)
Isolated execution			
Full five	59.03	0.00	0.00 (0.00 to 0.00)
Arrival first-three	47.22	21.53	31.35 (28.63 to 33.78)
Arrival stability	58.33	3.47	24.59 (23.14 to 26.23)
Quorum-3 else full	59.03	0.00	13.55 (12.05 to 15.06)
Completion-invariant	59.03	0.00	15.15 (13.32 to 16.99)
Shared execution			
Full five	59.03	0.00	0.00 (0.00 to 0.00)
Arrival first-three	48.26	21.88	26.58 (24.49 to 28.25)
Arrival stability	59.03	4.17	20.21 (18.55 to 21.84)
Quorum-3 else full	59.03	0.00	11.59 (9.89 to 13.30)
Completion-invariant	59.03	0.00	12.91 (10.90 to 14.93)

Arrival selection changes decisions beyond the mean quality of selected responses. Under shared execution, arrival first-three has 48.26% decision accuracy, compared with 59.03% for full five, and disagrees with full five on 21.88% of traces. Its selected-response accuracy shift is -0.97 points, with an interval from -3.56 to $+1.71$ points. This interval includes zero, whereas the final-decision contrast is negative. The distinction follows from the nonlinear plurality rule in Equation 5.

The execution condition changes participation even when every response stays fixed. Llama 3.1 8B has the highest model accuracy and enters 28.82% of isolated first-three sets and 36.11% of shared first-three sets. Table 24 gives all identities. Changing execution condition changes the first-three decision on five traces belonging to three tasks; its accuracy rises from 47.22% to 48.26%. Since every paired output is identical, these decision changes arise from the measured arrival order. The comparison of first-three with full five also changes vote count and is a separate contrast.

Table 24: Model accuracy and first-three inclusion under paired execution. Outputs are identical across conditions; inclusion can still change. Equal inclusion would be 60% for every identity.

Model	Accuracy (%)	Isolated inclusion (%)	Shared inclusion (%)
Qwen2.5 0.5B	25.69	66.67	66.67
Qwen2.5 1.5B	51.39	84.38	82.99
Qwen3 0.6B	47.92	54.51	51.39
Gemma 2 2B	47.22	65.62	62.85
Llama 3.1 8B	58.33	28.82	36.11

G.4 WORKLOADS, TASK GROUPS, AND ANSWER ENDPOINTS

Decision-preserving time reductions recur across the six workloads. Completion-invariant commit reduces shared submission time by 8.81% to 17.00% across workloads, with zero full-decision disagreement throughout. Arrival first-three has a negative full-decision accuracy contrast in every workload, whereas its selected-response shift has both signs (Table 25). This pattern supports reporting response composition and final decisions separately.

Table 25: Shared-execution results by workload. Each workload contains 24 tasks with two technical repetitions. Selected-response and decision shifts compare arrival first-three with full five. Completion-invariant commit has zero full-decision disagreement in every workload.

Workload	Selected shift (pp)	Decision shift (pp)	Exact commit reduction (%)
1	3.89	-4.17	13.89
2	-3.75	-4.17	8.81
3	-1.94	-25.00	12.65
4	3.33	-2.08	11.17
5	-2.78	-8.33	17.00
6	-4.58	-20.83	13.93

The tasks excluded from the deployment pilot retain the same decision-preserving benefit. On these 72 tasks, completion invariance reduces shared commit time by 11.70%, with a 95% workload-cluster interval from 9.63% to 14.41%, while preserving every full decision. Arrival first-three loses 9.03 accuracy points relative to full five, with an interval from 4.17 to 13.89 points. Table 26 reports both task groups and all three benchmarks. These two partitions reuse the same 144 tasks.

Table 26: Shared-execution results by benchmark and deployment-pilot membership. Decision shift compares arrival first-three with full five; commit reduction uses completion invariance. Parentheses give 95% workload-cluster bootstrap intervals. The benchmark rows and pilot-membership rows are two overlapping partitions of the same 144 tasks.

Group	Tasks	Decision shift (pp)	Exact commit reduction (%)
GSM8K	48	-11.46 (-25.00 to 0.00)	10.07 (7.87 to 12.97)
SVAMP	48	-18.75 (-29.17 to -8.33)	22.38 (18.93 to 26.38)
MATH-500 numeric	48	-2.08 (-6.25 to 0.00)	6.26 (4.33 to 8.20)
Pilot tasks	72	-12.50 (-26.39 to -0.00)	14.11 (10.65 to 16.42)
Nonpilot tasks	72	-9.03 (-13.89 to -4.17)	11.70 (9.63 to 14.41)

The endpoint protocol remains part of the aggregation target. In each execution condition, 1,006 of 1,440 responses have valid answer endpoints, and 98 of 288 full-panel decisions are abstentions. Completion invariance preserves these decisions as well as numeric answers. A descriptive split by the full decision gives a mean shared commit reduction of 17.80% on the 190 numeric-answer traces and 3.41% on the 98 abstention traces. Earlier numeric answers therefore contribute directly to the overall time reduction. The primary comparisons retain every task and the fixed 256-token budget.

H Reproducibility details

H.1 Assumptions and empirical scope

Table 27 separates the conditions required by each mathematical statement from the empirical setting used to evaluate it. The exact identities apply under their stated conditions, while the measured effect sizes describe the two evaluated panels.

H.2 Artifact inventory

The supplementary artifact is organized by scientific role. It contains experiment source, raw generations, task records, model-level metrics, trace-level policy replays, exact subset frontiers, task-cluster estimates, parser comparisons, runtime mappings, controlled ingress interventions, analysis code, figure-generation code, and environment manifests. Table 28 maps each core artifact to its scientific role.

Table 27: Conditions and empirical scope of the main mathematical objects.

Object	Condition for the stated result	Empirical instantiation and boundary
Selected-response law	Complete joint law of terminal marks and times; fixed selected-set size for Equation 3	Five completed shadow workers; variable-size policies use the separate laws in Appendix B
Identity decomposition	Integrable response feature and observed inclusion indicators	Correctness feature on complete five-model traces; it is an exact finite-sample identity
Joint decision law	Policy subset kernel and aggregator defined on the complete potential trace	Fixed plurality and common-trace policy replay; first-order identity weights alone do not determine decisions
Delay budget	Static first-k, nonnegative post-generation delay, and a fixed mark-independent tie rule	Controlled ingress replay with fixed payloads; shared scheduling, cancellation, and adaptive public-network control are outside this experiment
Completion invariance	The admissible completion set contains the realized unseen continuation	Exact plurality enumeration for five identities; no learned predictor or approximate stopping score is used
Effect estimates	A specified panel, runtime mapping, task mixture, and estimand	Two disjoint five-model panels on seven benchmarks, plus paired shared execution of the reasoning panel

Table 28: Core artifact files and their scientific role.

Artifact group	Contents and scientific role
Generation evidence	Raw outputs, task records, normalized marks, correctness, and service times
Policy evidence	Complete traces, subset policies, first-k paths, runtime mappings, and all 120 identity priorities
Statistical evidence	Model metrics, task-cluster estimates, parser bounds, decompositions, and bootstrap intervals
Shared execution	Paired completions, workload allocation, effective decoder settings, and workload-cluster estimates
Delay evidence	Target frontiers, constructive schedules, and verified ingress interventions
Reproduction code	Policy analysis, exact decomposition, tables, and deterministic raster figures

H.3 REPRODUCTION SEQUENCE

The reproduction cell downloads the public inputs and reruns the empirical sequence. Greedy outputs and measured times depend on the execution configuration, while analysis of a fixed complete trace is deterministic.

1. Load the specified public model releases and sampled benchmark records for each panel.
2. Generate both greedy executions and retain terminal text, normalized marks, correctness, and per-sequence service times.
3. Assemble five-model traces and apply full-five, arrival first-three, every preselected subset, and both decision-equivalent policies.
4. Average technical executions within tasks and bootstrap task clusters.
5. Apply the frozen multiple-choice parser, unresolved-answer bounds, and runtime mappings to the held-out panel.
6. Enumerate the reasoning-panel first-k path, all fixed identity priorities, and delay targets.
7. Replay the specified interventions through the ingress controller.
8. Execute the paired shared-GPU workloads and compute workload-cluster summaries with both repetitions retained.
9. Run the analysis and figure scripts directly from the packaged trace tables.

The full theory can be checked independently of model execution. Equations 3, 4, and 5 are conditioning identities. Equation 6 can be verified by enumerating all target subsets of a finite trace. Equation 7 can be verified by enumerating admissible unseen label-equality patterns and comparing each early commit with the full decision.

H.4 EVIDENCE SUMMARY

Summary of mathematical statements and corresponding experiments.

Claim	Mathematical object	Empirical evidence
Arrival changes effective identity weight	p_i/k in Equation 3	Ten identity inclusion propensities across two panels
Identity reweighting dominates item shift	Equation 4	Exact decompositions in both panels and four held-out mappings
Fewer votes and arrival selection are distinct	Equation 5	Exact subsets, first-k, and 120 tie priorities
Delay alone can change the decision	Equation 6	128 fixed-payload replays
Exact early commit preserves the full result	Equation 7	Zero mismatches on 672 original traces and 576 paired-execution traces