WHO IS WORTH WAITING FOR? DECISION CLOSURE IN HETEROGENEOUS LANGUAGE MODEL ENSEMBLES

Anonymous authors
Paper under double-blind review

ABSTRACT

A heterogeneous language model ensemble can reach a fixed decision before all its models finish: individual outputs may remain uncertain even when every compatible combination gives the same answer. We use this condition to stop unfinished calls while preserving the complete-panel decision. Split conformal prediction converts native token prefixes into nested sets of possible terminal outputs. The controller stops when these sets imply a unique decision and, before then, cancels individual calls that can no longer change it. We prove reference-decision fidelity on the joint coverage event for any fixed deterministic rule over finite outputs. Experiments with ten open models on six public benchmarks show wall-time reductions of 34.9% to 53.6% in paired low-load trials. On two complete science-question test splits totaling 1,672 questions, both panels save 49.0% to 71.0% of generated tokens and retain both reference decisions on 99.8% of questions. Weighted voting and strict majority with deferral also benefit from the same stopping condition.

1 INTRODUCTION

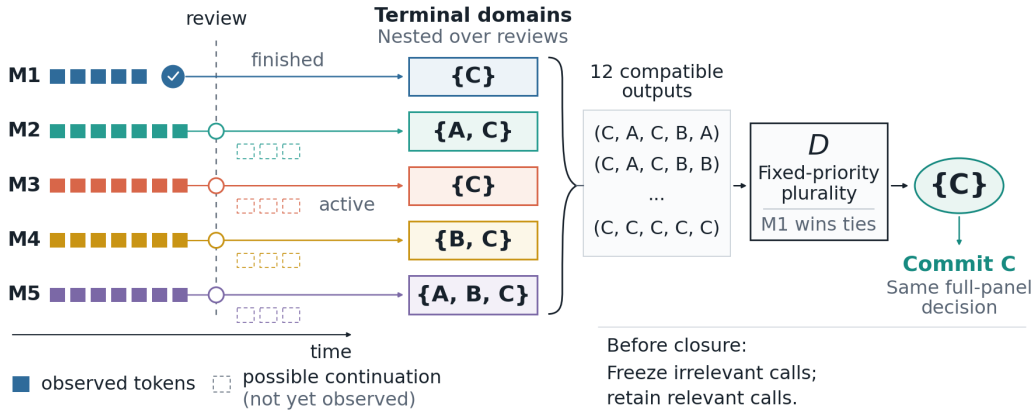


Figure 1: Decision closure before all models finish (illustrative example). M1 has returned C; the other terminal labels remain constrained by their prefixes. Fixed-priority plurality, with M1 first in ties, returns C for every compatible tuple. The remaining calls can therefore stop. Before closure, only irrelevant calls are frozen. Fidelity holds when the domains contain the terminal outputs.

Test-time computation can improve language-model decisions through multiple reasoning paths, models, or agents (Wang et al., 2023; Yao et al., 2023; Du et al., 2024; Li et al., 2024a; Wang et al., 2024; Snell et al., 2024). Parallel execution reduces waiting, yet some calls may continue after their joint prefixes determine the complete-panel decision. Stopping at that point could save time and tokens. The challenge is to recognize it without observing the unfinished outputs.

Efficiency methods act at several stages of model execution. Routers select a model before generation, and cascades decide whether to call another model after a complete response (Chen et al., 2024; Ong

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.

et al., 2024; Dekoninck et al., 2024; Gupta et al., 2024). Parallel stopping methods inspect partial reasoning traces and decide when to stop a homogeneous sample pool (Zheng et al., 2026; Chen et al., 2026). Serving systems schedule active computations (Lin et al., 2024; Luo et al., 2025; Huang et al., 2026b). We study the decision condition needed for a heterogeneous panel: which unfinished outputs can still alter its fixed full-computation decision, given the prefixes already observed?

Decision relevance depends on what the active streams jointly allow. An accurate or highly weighted model may become irrelevant once the other streams constrain the outcome. Its terminal output matters only if two values still allowed by its current stream lead to different decisions for some allowed assignment of the other outputs. A static model ranking cannot capture this dependence on the current state. Figure 1 illustrates the stopping condition.

We formalize this interface with set-valued terminal uncertainty. Each active computation i has a finite terminal space $\mathcal{Y}_i$ and a domain Γ_i^k at review epoch k. A fixed deterministic decision map D sends the terminal tuple to an action. We enumerate the image of the joint domain under D. The decision is closed when this image is a singleton. Coordinate i is relevant when changing only its allowed value can alter the decision for some compatible tuple. Domain contraction makes irrelevance persistent, so irrelevant calls may be frozen together. This statement is pathwise and applies to any fixed finite deterministic map.

The remaining issue is how to obtain useful domains from language-model streams. We use the answer mark parsed from each native token prefix and fit a categorical terminal-mark model on disjoint development tasks. A source-wise split-conformal score covers every registered review state of a task as one group. Raw prediction sets are intersected over time, with the preceding set retained when an intersection is empty. This construction supplies nested domains and a finite-sample risk interface under source-wise exchangeability. It neither assumes monotone prefix answers nor requires extra answer-forcing completions.

The evidence tests the decision condition, its resource benefit, and the source of that benefit. A general theorem proves fidelity on the joint coverage event, and exhaustive finite-state checks test contraction and freezing. An initial study of two five-model panels on six benchmarks preserves every reference decision. Its fixed-trace decomposition shows that completed votes remove 28.3% of tokens, prefix domains raise the reduction to 42.0%, and freezing raises it to 44.0%. Further studies add 288 questions, a deferral rule, and complete tests for OpenBookQA and the Challenge subset of the AI2 Reasoning Challenge (ARC-Challenge).

The analysis and experiments address complementary questions. The closure theorem establishes when stopping preserves a fixed decision; conformal calibration controls the risk that the domains miss the terminal outputs. Randomized live trials measure resource savings, while matched traces separate the contribution of prefix information from that of cancelling individual calls.

2 Related Work

Additional test-time computation improves reasoning when diverse samples or agents correct one another. Self-consistency aggregates sampled chains (Wang et al., 2023), Tree of Thoughts searches over intermediate states (Yao et al., 2023), and multiagent debate shares evidence among model instances (Du et al., 2024). Model and agent ensembles aggregate responses from larger groups of model instances (Li et al., 2024a; Wang et al., 2024). These methods motivate launching several computations. Our question begins after launch: when has their fixed downstream decision become determined?

Adaptive allocation reduces the cost of invoking multiple models. FrugalGPT, AutoMix, and RouteLLM learn query-dependent choices among heterogeneous models (Chen et al., 2024; Aggarwal et al., 2024; Ong et al., 2024). RouterBench provides common data for evaluating such policies (Hu et al., 2024). Language-model cascades and unified routing formulations decide whether a completed response justifies another call (Gupta et al., 2024; Dekoninck et al., 2024). Recent controllers adapt model selection after complete responses or during multi-step workflows (Huang et al., 2026a; Li et al., 2026). Judge-panel design instead allocates evaluators from audit data and declared slices (Zhu et al., 2026). These actions concern computations that have not begun. Our action concerns a call that is already consuming resources, so only its remaining generation can be saved.

Parallel stopping moves observation inside unfinished computation. Early-stopping self-consistency, certified self-consistency, and adaptive ensemble consensus decide from complete samples or deliberation rounds (Li et al., 2024b; Cordero-Encinar & Duncan, 2025; Medina, 2026). Parallel-Probe reads two-dimensional partial reasoning states and uses consensus stability to stop or prune branches (Zheng et al., 2026). Margin-Adversarial Risk-controlled Stopping (MARS) calibrates a margin-based risk criterion for global stopping in parallel test-time scaling (Chen et al., 2026). These works establish that partial traces can support early decisions. We derive exact coordinate relevance for distinct active models under a fixed downstream map, which yields global closure and selective freezing from the same decision image.

Agent orchestration and serving systems address the physical execution of model graphs. Aegean uses progressive agreement and straggler cancellation, while latency-aware orchestration controls critical paths and future interactions (Ruan et al., 2025; Shi et al., 2026). Parrot, Autellix, SwarmX, and orchestration across heterogeneous accelerators optimize semantic dependencies, program-level scheduling, tail latency, and physical placement (Lin et al., 2024; Luo et al., 2025; Huang et al., 2026b; Wang et al., 2026). They build on work on tail latency (Dean & Barroso, 2013) and memory-efficient model serving (Kwon et al., 2023). Our controller tests whether an active computation is still required by a fixed decision. A serving scheduler can execute that test efficiently, but faster scheduling alone cannot derive it.

Conformal prediction provides distribution-free coverage under exchangeability (Vovk et al., 2005; Angelopoulos & Bates, 2023). Adaptive prediction sets have been developed for classification (Romano et al., 2020; Angelopoulos et al., 2021). We use split conformal prediction at the task-trajectory level: one calibration score covers all model identities and review epochs registered for a task. The contribution lies in propagating these sets through the downstream decision map and preserving coordinate irrelevance across sequential reviews. Appendix B compares the observation, action, and fidelity interfaces of the closest controllers.

3 Post-Launch Decision Closure

3.1 Setting

Consider m computations launched in parallel for an input x. Computation i ends with a value $Y_i \in \mathcal{Y}_i$, where every $\mathcal{Y}_i$ is finite. The downstream system has a fixed deterministic map

$$D : \prod_{i=1}^{m} \mathcal{Y}_i \to \mathcal{A}, \tag{1}$$

where $\mathcal{A}$ is the action space. The reference action is $D(Y_1, \ldots, Y_m)$ after all computations finish. Our objective is to reproduce this action earlier. We assume that cancellation preserves the terminal values of retained calls; the live comparisons check this condition for the evaluated system. The aggregator remains fixed.

Review epochs $k = 1, 2, \ldots$ occur when a call finishes or an external timer requests a review. At epoch k, the system observes every completed value and the current native prefix of every active call. It maintains a nonempty terminal domain $\Gamma_i^k \subseteq \mathcal{Y}_i$. A completed value has the singleton domain $\{Y_i\}$. An active domain may contract as more tokens arrive. If a call is cancelled, its current domain is frozen.

3.2 Decision image and coordinate relevance

The joint domain determines every action still compatible with the current state.

Definition 1 (Decision image)**.** *At review epoch k, the set of possible full-computation actions is*

$$\mathfrak{D}_k = \left\{ D(y_1, \ldots, y_m) : y_i \in \Gamma_i^k \right\}. \tag{2}$$

The decision is closed when $|\mathfrak{D}_k| = 1$.

Closure is global. Before it occurs, a coordinate test identifies which individual calls can still alter the downstream action.

Definition 2 (Decision relevance)**.** *Let y_{-i} denote an assignment to every coordinate except i. Coordinate i is relevant at epoch k if there exist $a, b \in \Gamma_i^k$ and $y_{-i} \in \prod_{j \neq i} \Gamma_j^k$ such that*

$$D(y_{-i}, a) \neq D(y_{-i}, b). \tag{3}$$

We write R_k for the set of relevant active coordinates.

This definition uses the current joint domain. It can therefore mark a strong model irrelevant after the other calls constrain the action, or retain a small model whose remaining values still separate two actions. The controller commits the unique element of $\mathfrak{D}_k$ when closed. Otherwise, it retains R_k and freezes every active coordinate outside R_k.

3.3 Nested terminal domains from token prefixes

For the large language model (LLM) instance, $\mathcal{Y}_i$ contains the valid multiple-choice labels and an abstention symbol $\perp$. Let M_i^k be the canonical mark parsed from model i's prefix at epoch k, and let L be the number of valid labels. On independent fit tasks, counts $N(m, L, y)$ define the smoothed categorical estimate

$$\widehat{p}(y \mid m, L) = \frac{N(m, L, y) + 1}{\sum_{y'} N(m, L, y') + L + 1}. \tag{4}$$

The estimate is pooled across model identities. Identity still enters the fixed aggregation rule and the coordinate test in Equation 3.

Calibration treats a complete task trajectory as one exchangeable group. For calibration task g, let $\mathcal{O}_g$ contain every registered pending identity and review epoch across both panels and loads. Its score is

$$S_g = \max_{(k,i) \in \mathcal{O}_g} \left[1 - \widehat{p}(Y_i \mid M_i^k, L)\right]. \tag{5}$$

For benchmark source s with n_s calibration groups, let q_s be the $\lceil (n_s + 1)(1 - \alpha) \rceil$th ordered score, or $+\infty$ if this rank exceeds n_s. This finite-sample quantile gives the raw domain

$$C_i^k = \left\{ y \in \mathcal{Y}_i : 1 - \widehat{p}(y \mid M_i^k, L) \leq q_s \right\}. \tag{6}$$

An empty raw domain is replaced by $\mathcal{Y}_i$. Taking a maximum in Equation 5 covers the entire registered trajectory, including the random epoch at which the controller stops.

Raw conformal sets need not contract between reviews. That property is essential because a coordinate proved irrelevant under a smaller domain can become relevant if the domain later expands. We therefore update active domains by

$$\Gamma_i^k = \begin{cases} \Gamma_i^{k-1} \cap C_i^k, & \text{if the intersection is nonempty,} \\ \Gamma_i^{k-1}, & \text{otherwise,} \end{cases} \tag{7}$$

with $\Gamma_i^1 = C_i^1$. Thus $\Gamma_i^k \subseteq \Gamma_i^{k-1}$ by construction. Appendix D gives a two-coordinate example showing why this condition cannot be removed.

3.4 Controller implementation

The controller has one repeated operation. At a review epoch, it reads current prefixes, updates active domains with Equation 7, inserts exact singleton domains for completed calls, and computes $\mathfrak{D}_k$. A singleton image is returned immediately. Otherwise, Equation 3 is evaluated for each active coordinate, irrelevant calls are cancelled, and their domains are frozen. The next review occurs at a retained completion or an external review time.

Direct enumeration costs $O(|\prod_i \Gamma_i^k|)$ evaluations of D. This is small for the five-model multiple-choice panels used here. Structured decision maps can admit cheaper tests, but the theorem does not require such structure. Appendix E gives pseudocode, complexity details, and the fixed-priority plurality specialization.

4 GUARANTEES

The guarantee separates a deterministic pathwise statement from the statistical event that supplies its domains.

Theorem 1 (Decision closure under nested domains)**.** *Fix a terminal tuple $y^* \in \prod_i \mathcal{Y}_i$. Suppose every visited domain contains its terminal value, active domains are nested, and cancelled domains remain frozen. Then any coordinate outside R_k may be frozen without changing the decision image reachable from the current domains. If $\mathfrak{D}_k = \{d\}$ at a later epoch, then $d = D(y^*)$.*

The proof follows by coordinate replacement. If $i \notin R_k$, changing coordinate i between any two values in Γ_i^k leaves D unchanged for every fixed assignment of the other coordinates. Repeating the replacement freezes any collection of irrelevant coordinates. Later contraction cannot create a witness that was absent from the larger domain. Since y^* remains in the active or frozen product, a singleton image must contain $D(y^*)$. Appendix C gives the complete proof and the simultaneous-freezing lemma.

For source s, let the fit data be independent and let its n_s calibration task groups and a test group be exchangeable. The split-conformal rank argument applied to Equation 5 gives

$$\Pr\big[Y_i \in C_i^k \text{ for every registered } (k,i)\big] \geq 1 - \alpha. \tag{8}$$

On this event, every nonempty intersection in Equation 7 contains the true value. Theorem 1 therefore bounds source-wise decision disagreement with the same full-computation tuple by α. This is a risk guarantee for reference-decision fidelity. It does not imply lower latency, which must be measured under a live intervention.

The theorem is not tied to plurality or language models. We exhaustively checked all 256 binary-output maps on three binary coordinates and all 19,683 ternary-output maps on two ternary coordinates. Across 5,022,443 singleton fidelity checks and 3,220,638 nested irrelevance checks, no violation occurred. These computations test the finite implementation of the set logic. The proof remains the source of generality.

5 EXPERIMENTAL DESIGN

The experiments test reference fidelity, live resource savings, and their mechanism. The initial protocol fixes both panels, six sources, and two loads before test generation. Extensions add 288 disjoint questions with the frozen controller and the complete OpenBookQA and ARC-Challenge test splits with validation-only fitting and calibration (Appendices I, J, and K).

5.1 MODELS, TASKS, AND REFERENCE DECISION

We form two disjoint heterogeneous panels. Panel A contains OLMo-2-1B, SmolLM2-1.7B, Granite-3.3-2B, Phi-3.5-mini, and Mistral-7B-Instruct-v0.3. Panel B contains Qwen2.5-3B, Qwen3-4B, Gemma-2-2B, Phi-4-mini, and OLMo-2-7B. The models range from one to seven billion parameters and come from seven model families.

The confirmation set has 96 fresh questions, with 16 from each of ARC-Challenge, HellaSwag, Massive Multitask Language Understanding (MMLU), TruthfulQA multiple-choice 1 (MC1), CommonsenseQA, and OpenBookQA (Clark et al., 2018; Zellers et al., 2019; Hendrycks et al., 2021; Lin et al., 2022; Talmor et al., 2019; Mihaylov et al., 2018). These identities do not overlap any fit, calibration, or method-development task. We use greedy decoding with at most 128 generated tokens. The parser accepts an explicit final label or a leading answer label. Other text maps to $\perp$.

The reference map is abstention-neutral fixed-priority plurality. Invalid and abstaining marks cast no vote. Ties are resolved by the highest-priority model among the tied labels, and the first valid label is the deterministic fallback when no valid vote exists. Each panel's priority order is fixed before testing. The method is evaluated against this reference, so equal controller and full-panel accuracy is the expected fidelity outcome.

5.2 Calibration and live arms

In the initial study, six disjoint tasks per source fit Equation 4 and 20 calibrate Equation 5. We set $\alpha = 0.05$, which selects the twentieth ordered score from each source. All panel identities, loads, and review epochs of a task enter one trajectory-level score. No confirmation task updates the estimate or quantile.

Two physical arms isolate real resource use. The full-trace arm launches all five models and waits for every output. The closure arm launches the same calls, reviews all active prefixes at natural completion events, and applies the controller. Low load runs one query at a time and uses task-level crossover with balanced arm order. The batched condition executes two 48-task workloads per panel with opposite arm orders after throughput-based capacity calibration. These measurements provide a directional stress test within one runtime session.

The primary fidelity unit is the task group: both panels, both loads, and all review states. We report task-level disagreement and raw simultaneous-coverage failures with one-sided 95% Clopper–Pearson upper bounds (Clopper & Pearson, 1934). Same-trace fidelity compares closure with full plurality on the uninterrupted reference trace. Live fidelity compares the cancelled execution with the separately reconstructed shadow action. For low-load resources, 10,000-draw paired bootstrap intervals resample task identities within each source and panel. Appendix F specifies randomization, estimands, prompts, parsing, and execution.

6 Results

6.1 The controller preserves the full-panel decision

The initial study gives exact agreement on all 96 questions. No task produced a controller decision that differed from full-panel plurality, giving a one-sided 95% upper bound of 3.07%. Same-trace fidelity and live-to-shadow fidelity were both 384 of 384 trajectories. Controller and reference accuracy were therefore identical at 70.31%.

Table 1: Confirmation reference-decision fidelity. A task group contains both panels and loads. Bounds are one-sided 95% Clopper–Pearson limits.

Endpoint	Observation	Failure upper bound
Task decision disagreement	0/96	3.07%
Task raw coverage failure	1/96	4.85%
Same-trace decision fidelity	384/384	Task-level bound
Live-to-shadow decision fidelity	384/384	Task-level bound
Closure / full accuracy	70.31% / 70.31%	Not applicable

Decision fidelity held even on the single task with a raw coverage miss. Raw simultaneous coverage failed on one OpenBookQA task; the pooled binomial upper bound is 4.85%. This pooled summary is distinct from the source-wise conformal guarantee. The missed model prefix changed from label D to label A after its raw domain had become $\{D\}$. That coordinate did not change the aggregate decision. The case illustrates the theorem's implication: joint coverage is sufficient for fidelity, while a coverage miss need not produce a decision error.

The controller also closed before full completion in most trajectories. It committed before the fifth completion in 291 of 384 traces and observed fewer than five completed calls in 331. The closure index, defined as the number of completed calls when an uninterrupted trace first closes, averaged 2.99. Thus perfect agreement did not arise from routinely waiting for the full panel.

Closure recurs in two distinct panel regimes. Panel A has 64.06% full-panel accuracy and a mean closure index of 2.76, whereas Panel B reaches 76.56% accuracy and closes at 3.22. The controller preserves every reference decision in both panels. Their different closure times also predict the resource results: the panel that settles earlier leaves more unfinished generation to remove. This comparison separates decision fidelity, which holds in both regimes, from the amount of available saving.

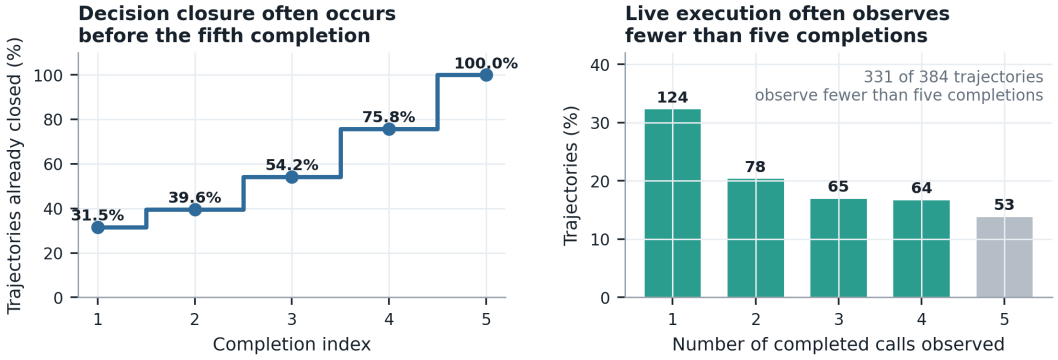


Figure 2: Most fresh trajectories close before complete-panel execution. The left panel gives the cumulative closure distribution on uninterrupted traces. The right panel counts completion events observed by the live controller after coordinate freezing. Counts sum to 384 trajectories.

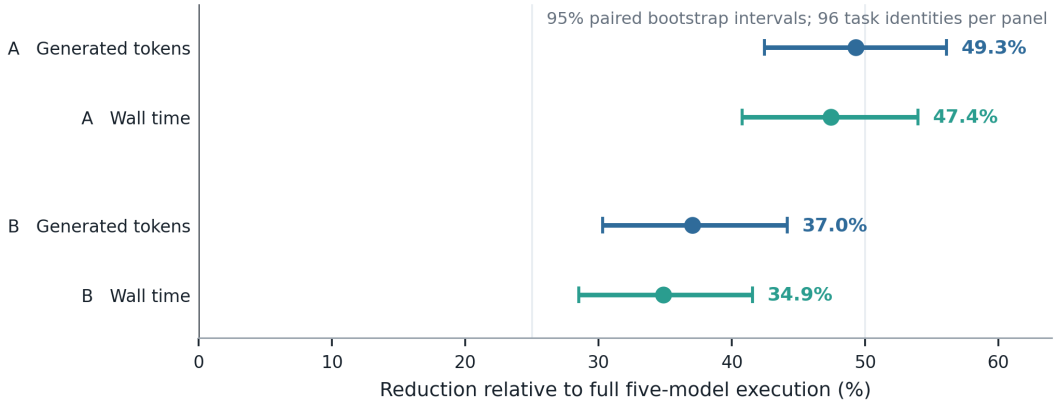


Figure 3: The complete controller reduces both live generated work and wall time under low load. Points show reductions relative to full five-model execution. Bars are 95% paired bootstrap intervals over 96 task identities per panel. Accuracy and the fixed full-panel decision are unchanged.

6.2 Live token and wall-time savings

Under low load, early stopping reduces both generated tokens and elapsed time. Figure 3 reports paired reductions with bootstrap intervals over task identities. Panel A generated 75.5 tokens per task, compared with 149.0 under full execution, and released the arm in 0.84 seconds, compared with 1.59. These changes reduce tokens by 49.3% and wall time by 47.4%. Panel B generated 113.1 tokens, compared with 179.6, and released the arm in 1.50 seconds, compared with 2.30, for reductions of 37.0% and 34.9%. All four intervals exclude zero.

Batched execution preserves reductions in every panel and endpoint. Generated tokens fell by 50.0% in Panel A and 36.3% in Panel B. Wall time fell by 9.3% and 3.3%, respectively, and both opposite arm-order clusters favored the controller in every panel-metric pair. Because each panel contributes two workload pairs within one session, this study tests directional replication rather than a precise high-load effect. We therefore base quantitative inference on the low-load paired comparison.

Both panels retain their token savings under batching. Table 2 shows 100% decision fidelity in each panel and load. Generated-token reductions are stable across loads within a panel, while wall-time reductions narrow under batching because retained and cancelled work shares the same batched execution path. At the benchmark level, all 24 source-panel resource comparisons favor closure. The detailed source effects appear in Appendix G.

Table 2: Decision fidelity and resource reductions. Low load uses 96 paired tasks; batched rows use two oppositely ordered 48-task workloads.

Panel	Load	Trajectories	Fidelity	Token saving	Time saving
A	Low	96	100%	49.3%	47.4%
A	Batched	96	100%	50.0%	9.3%
B	Low	96	100%	37.0%	34.9%
B	Batched	96	100%	36.3%	3.3%

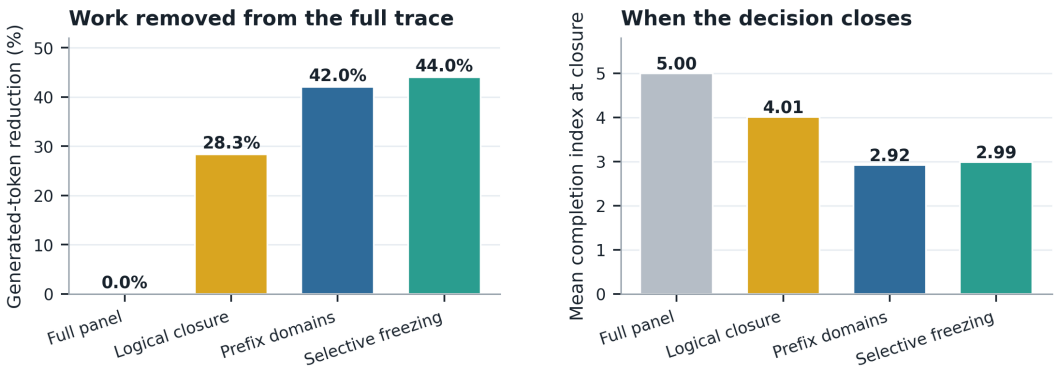


Figure 4: Native prefix domains supply the main gain beyond completion-only logical closure. All four policies are evaluated on the same 384 uninterrupted traces and preserve the full-panel decision. Generated-token reductions are shadow quantities in this decomposition. The live effects appear in Figure 3.

The extension reproduces low-load savings on 48 new questions per panel. Tokens fall by 56.8% and 41.3%, and wall time by 53.6% and 39.0%. Across all 288 new questions, reference fidelity holds on 287. Prefix-informed closure also benefits strict-majority voting with deferral (Appendix I).

Two complete official test splits extend the result beyond benchmark subsets. With validation-only fitting and calibration, both panel decisions are preserved on 499 of 500 OpenBookQA questions and 1,170 of 1,172 ARC-Challenge questions. Across the four benchmark-panel combinations, generated tokens fall by 49.0% to 71.0% and paired workload time by 4.3% to 25.0%. Total correct-answer counts are unchanged within each benchmark. Appendices J and K report panel-level accuracy, history groups, every workload pair, and all reference disagreements.

With validation-fitted voting weights, fixed-trace replay saves 52.6% to 68.4% of tokens and preserves both weighted references on 1,670 of 1,672 questions (Appendix L).

6.3 Prefix domains add savings beyond completed votes

The mechanism decomposition separates three sources of saved work on the 384 uninterrupted initial traces. Logical closure with unrestricted pending domains removes 28.3% of generated tokens because completed votes can sometimes determine plurality. Replacing unrestricted domains with calibrated prefix domains raises the reduction to 42.0%. Selective coordinate freezing raises it to 44.0%. Figure 4 shows the same ordering and the mean completion index at closure.

Coordinate freezing saves work without guaranteeing an earlier review. Relative to prefix-domain global closure, it removes another 3.23 tokens per trajectory and avoids 4,259 precommit token steps in total. The closure index is unchanged on 362 trajectories and later on 22, with no earlier cases. A cancelled call may have been irrelevant to the terminal decision while its completion event would have triggered an earlier read of the other active prefixes. This distinction motivates the discussion that follows.

The additional saving from prefixes persists under task-level resampling. A paired bootstrap, stratified by source and retaining all four executions of each task, estimates an extra 13.7 percentage points

of token reduction over logical closure, with a 95% interval from 11.1 to 16.4 points. Prefix domains account for 87.5% of the improvement beyond completed votes. Appendix H reports the corresponding freezing contrast and source breakdown.

The decision rule also handles uninformative prefixes. Mean raw-set size ranges from 2.76 to 5.56 across sources. On HellaSwag, every pending domain retains all four labels and abstention, so its saving comes entirely from completed votes. On the other five sources, prefix domains provide additional token reductions. The same decision-image test therefore uses prefix information where it constrains the outcome and otherwise reduces to logical closure.

7 Discussion

After launch, the value of a call depends on the other active streams. Routing models estimate which system to call, often from the query alone (Chen et al., 2024; Ong et al., 2024; Hu et al., 2024). Once calls are active, the relevant object is the image of their joint uncertainty under the downstream decision. This image can change after any completion or informative prefix update. Model quality and speed remain useful for dispatch, but neither establishes whether a running call can still affect the decision.

These roles support a modular implementation. Routers and cascades allocate calls (Gupta et al., 2024; Dekoninck et al., 2024), our controller tests active streams, and serving systems execute retention or cancellation (Lin et al., 2024; Luo et al., 2025; Huang et al., 2026b). The interfaces share a declared decision rule without requiring a shared internal score.

Decision closure gives global stopping and selective cancellation a common condition. Parallel-Probe combines consensus stopping with branch pruning; MARS controls the risk of changing the full-budget vote (Zheng et al., 2026; Chen et al., 2026). Our decision image tests a fixed rule over distinct models: a singleton permits commitment, and coordinate invariance permits freezing. In the initial study, global prefix-informed closure supplies most of the measured saving, with further token reduction from freezing.

Trajectory-level calibration matches the unit at which the controller adapts. Split conformal prediction calibrates one score per exchangeable example (Vovk et al., 2005; Angelopoulos & Bates, 2023). Here, that example is a task trajectory, with the maximum nonconformity over all registered models, loads, and reviews. This grouping covers the selected stopping state, while nested intersection supplies the contraction required by the relevance proof. The coverage event links the statistical and pathwise arguments.

Terminal decision relevance is narrower than the full value of waiting. A call whose terminal value cannot change D may still trigger a useful review of other prefixes. Removing that event can delay closure and offset saved generation, as observed for weighted voting in one source-panel combination. Latency-aware orchestration and serving systems can address this scheduling dimension (Ruan et al., 2025; Shi et al., 2026; Lin et al., 2024; Luo et al., 2025; Huang et al., 2026b; Wang et al., 2026). Combining semantic relevance with explicit review timing is therefore a distinct optimization problem.

The evaluated rules separate the decision target from the stopping condition. Weighted voting assigns unequal model contributions, while strict majority can defer. Both benefit from the same prefix domains used for plurality. The evidence spans two five-model panels and complete OpenBookQA and ARC-Challenge tests. Free-form use requires a finite decision representation and representative calibration. Independent sessions would quantify deployment variation beyond the paired workload savings measured here.

8 Conclusion

We derive closure and relevance tests for calls already running in parallel. Nested conformal domains identify when unfinished outputs can still change a fixed decision. Ten models on six benchmarks support low-load wall-time reductions of 34.9% to 53.6%, with batched savings on two complete official test splits. Fixed-trace tests also support weighted voting and deferral. An aggregate decision can justify stopping while individual outputs remain uncertain.

AI USE STATEMENT

Generative AI tools supported mathematical formalization, experimental code implementation and checking, and manuscript polishing. The authors developed the theory 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 work studies resource control for publicly available language models and public multiple-choice benchmarks. It uses no human subjects, private data, or model-generated personal information. Early cancellation can change resource allocation across model providers, so deployment should preserve the declared reference rule and audit coverage under the target distribution. The appendix reports all observed decision disagreements, coverage failures, and resource effects.

REPRODUCIBILITY STATEMENT

Appendix F specifies task splits, models, prompts, parsing, calibration, randomization, runtime arms, and statistical units. Appendices G, I, J, and K provide complete results, including coverage failures and high-load effects. Appendix L gives the weighted-rule replay and probability-threshold comparison. Appendix N maps claims to code and result artifacts. Complete proofs appear in Appendix C. The complete reproduction code and experimental results are provided in the Supplementary Material.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, Shyam Upadhyay, Manaal Faruqui, and Mausam. AutoMix: Automatically mixing language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2310.12963.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Anastasios N. Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eNdiU_DbM9.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=cSimKw5p6R.

Wenbo Chen, Puheng Li, Mengyang Liu, Weijie Su, and Tianpei Xie. MARS: Margin-adversarial risk-controlled stopping for parallel LLM test-time scaling. *arXiv preprint arXiv:2606.12935*, 2026. URL https://arxiv.org/abs/2606.12935.

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.

Charles J. Clopper and Egon S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Paula Cordero-Encinar and Andrew B. Duncan. Certified self-consistency: Statistical guarantees and test-time training for reliable reasoning in LLMs. *arXiv preprint arXiv:2510.17472*, 2025. URL https://arxiv.org/abs/2510.17472.

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

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for LLMs. *arXiv preprint arXiv:2410.10347*, 2024. URL https://arxiv.org/abs/2410.10347.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2305.14325.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KgaBScZ4VI.

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

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. RouterBench: A benchmark for multi-LLM routing system. *arXiv preprint arXiv:2403.12031*, 2024. URL https://arxiv.org/abs/2403.12031.

Shihong Huang, Shengjie Wang, Hong Ma, and Zhou Xu. RLCascadeRouter: Quality-estimator-free cascade routing via reinforcement learning. *arXiv preprint arXiv:2608.15817*, 2026a. URL https://arxiv.org/abs/2608.15817.

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*, 2026b. URL https://arxiv.org/abs/2606.21401.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *ACM Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

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

Songyuan Li, Ahmed M. Abdelmoniem, and Shiqiang Wang. ProgRouter: Online progress-guided orchestration for multi-agent LLM workflows under quality-cost tradeoffs. *arXiv preprint arXiv:2608.25992*, 2026. URL https://arxiv.org/abs/2608.25992.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. *arXiv preprint arXiv:2401.10480*, 2024b. URL https://arxiv.org/abs/2401.10480.

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 *USENIX Symposium on Operating Systems Design and Implementation*, pp. 929–945, 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 *Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.

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.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018. doi: 10.18653/v1/D18-1260.

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.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.02544.

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.

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.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. URL https://arxiv.org/abs/2408.03314.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4149–4158, 2019. doi: 10.18653/v1/N19-1421.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Jinghao Wang, Yifeng Zhang, Xiao Zhou, Yao Lu, Yihui Zhang, Xiaoyang Sun, Tianyu Wo, Xu Wang, Chunming Hu, and Renyu Yang. Latency-aware orchestration for multi-agent LLM workflows on heterogeneous GPUs. *arXiv preprint arXiv:2609.03335*, 2026. URL https://arxiv.org/abs/2609.03335.

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. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.10601.

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

Tong Zheng, Chengsong Huang, Runpeng Dai, Yun He, Rui Liu, Xin Ni, Huiwen Bao, Kaishen Wang, Hongtu Zhu, Jiaxin Huang, Furong Huang, and Heng Huang. Parallel-probe: Towards efficient parallel thinking via 2d probing. *arXiv preprint arXiv:2602.03845*, 2026. URL https://arxiv.org/abs/2602.03845.

Bin Zhu, Yi Xie, and Yanghui Rao. Stopping and routing LLM judge panels. *arXiv preprint arXiv:2608.19802*, 2026. URL https://arxiv.org/abs/2608.19802.

A Appendix Guide

This appendix provides complete proofs, protocols, results, and reproduction details for the main text. Table 3 gives the reading order. The mathematical sections establish the result for a fixed deterministic decision map. The empirical sections evaluate fixed-priority plurality and extend the same domain construction to weighted voting and strict majority with deferral. The initial study, new-question extension, and complete OpenBookQA and ARC-Challenge tests are reported separately because their workload and calibration protocols differ.

Table 3: Appendix sections and their connection to the main claims.

Section	Purpose
B	Compares the control time, observed state, action, and declared target of the closest routing, stopping, orchestration, and serving methods.
C	States the invariants, proves simultaneous freezing and decision closure, and derives the split-conformal decision-risk corollary.
D	Gives minimal counterexamples for nonnested domains and for equating terminal relevance with review-time value.
E	Specifies the controller, exact finite-domain computation, stopping conditions, and complexity.
F	Records models, public tasks, prompts, parsing, calibration, aggregation, randomization, loads, and estimands.
G	Reports complete fidelity, accuracy, resource, source, model, and execution results.
H	Decomposes logical closure, prefix-domain closure, and coordinate freezing on the same fresh traces.
I	Tests 288 new questions, repeated live workloads, and strict majority with deferral using the frozen domains.
J	Evaluates all 500 OpenBookQA test questions with validation-only calibration, including history groups, every workload pair, and fixed-rule replay.
K	Repeats full-test evaluation on all 1,172 ARC-Challenge questions, retaining three-, four-, and five-option questions and both history groups.
L	Tests validation-fitted weighted voting on both complete test splits and compares exact closure with a probability threshold on the same trajectories.
M	Reports raw-record checks, fixed-trace policy comparisons, finite-state enumeration, and statistical stress tests.
N	Maps each claim to the required artifact class and expected output.

B Control Interfaces in Prior Work

Existing controllers differ in when they act, what they observe, and what outcome they preserve. Table 4 organizes the closest work by these dimensions. Routing and cascades allocate calls before launch or between complete responses. Parallel stopping observes active reasoning pools, agent protocols act on agreement, and serving systems schedule declared workflows. Decision closure instead controls distinct calls that are already generating and evaluates both global commitment and coordinate relevance under a fixed downstream map.

C Proofs

C.1 State and invariants

The controller acts on terminal uncertainty, not on a forecast of model quality. Let $\mathcal{I} = \{1, \dots, m\}$ index computations, let $\mathcal{Y}_i$ be the finite terminal space of computation i, and let $D : \prod_i \mathcal{Y}_i \to \mathcal{A}$ be fixed before execution. At review epoch k, each coordinate has a nonempty domain $\Gamma_i^k \subseteq \mathcal{Y}_i$. A completed coordinate has domain $\{Y_i\}$. An active coordinate may contract, while a frozen coordinate retains the domain present when its computation stopped.

Four conditions separate the pathwise argument from the statistical construction. The decision map is fixed, pending-domain updates contract, and frozen domains remain unchanged. These three properties follow from the controller. The fourth condition is coverage: the full-computation terminal

Table 4: Control interfaces in the closest literature. The entries describe each method's declared problem and do not rank performance across different settings.

Interface	State and control time	Action	Declared target
Routers and cascades	Query before launch or complete responses between calls	Select, launch, or escalate a model	Response quality under resource cost
Complete-sample stopping	Completed answers or deliberation rounds	Stop drawing samples or commit	Ensemble decision
Parallel-Probe	Elicited intermediate answers during homogeneous parallel reasoning	Stop globally or prune branches	Accuracy and computation
MARS	Probed answers and vote margin during homogeneous parallel reasoning	Stop globally	Risk-controlled full-budget vote
Aegean	Agent agreement during iterative refinement	Terminate refinement early	Consensus safety and liveness
Serving and orchestration	Dependencies and runtime state during workflow execution	Schedule workflows or calls	Latency and cost under workflow semantics
Decision closure	Exact outcomes and prefix domains during heterogeneous active generation	Commit globally or freeze a coordinate	Risk-controlled fixed-map full-computation decision

value Y_i belongs to every domain used for coordinate i. On this event, inserting an exact value at completion also contracts the preceding domain. Outside coverage, that singleton can fall outside the preceding set; the pathwise guarantee then does not apply.

For a product domain $G = \prod_i G_i$, write

$$D(G) = \{D(y) : y \in G\}. \tag{9}$$

Coordinate i is irrelevant on G when, for every $y_{-i} \in \prod_{j \neq i} G_j$ and every $a, b \in G_i$, $D(y_{-i}, a) = D(y_{-i}, b)$. This is the negation of Equation 3.

C.2 Simultaneous coordinate freezing

The first result shows that individually irrelevant coordinates can be stopped together. This matters because the implementation issues all currently justified cancellation requests at one review epoch.

Proposition 1 (Simultaneous freezing)**.** *Let $G = \prod_i G_i$, and let J be a set of coordinates that are each irrelevant on G. Choose any representative $r_i \in G_i$ for every $i \in J$. Then*

$$D(G) = D\left(\prod_{i \notin J} G_i \times \prod_{i \in J} \{r_i\}\right). \tag{10}$$

The same equality holds if each G_i, $i \in J$, is retained as a frozen set instead of being replaced by a representative.

Proof. The right-hand product is a subset of G, so its image is contained in $D(G)$. For the reverse inclusion, take any $y \in G$. Replace the coordinates in J one at a time by their representatives. At each replacement, all other coordinates still belong to their original domains. Irrelevance therefore implies that the value of D does not change. The resulting tuple belongs to the right-hand product and has decision $D(y)$. Every element of $D(G)$ is consequently present in the restricted image. Retaining the original set as a frozen set leaves the product, and hence its image, unchanged. □

The next result explains why an irrelevant computation need not be restarted after later reviews.

Proposition 2 (Persistence under contraction)**.** *Suppose coordinate i is irrelevant on $G = \prod_j G_j$. If $G'_j \subseteq G_j$ for every j, then coordinate i is irrelevant on $G' = \prod_j G'_j$.*

Proof. Any witness of relevance on G' would consist of $a, b \in G'_i$ and $y_{-i} \in \prod_{j \neq i} G'_j$ with different decisions. Because every G'_j is a subset of G_j, the same tuple would witness relevance on G, contradicting the premise. □

C.3 Proof of decision closure

We now prove Theorem 1 in full. The statement permits either representation of a stopped coordinate: its whole frozen domain or an arbitrary representative from that domain.

Proof of Theorem 1. Fix a visited epoch k. Proposition 1 shows that all coordinates outside R_k can be frozen simultaneously without changing the current decision image. Every later active domain is a subset of its domain at epoch k. Proposition 2 therefore prevents a frozen coordinate from becoming relevant after later contractions.

Consider any later epoch ℓ at which the controller commits. On the stated coverage event, the full-computation tuple y^* belongs to the product of all current active domains and all frozen domains. Hence $D(y^*) \in \mathfrak{D}_\ell$. Closure gives $\mathfrak{D}_\ell = \{d\}$, so $d = D(y^*)$.

If a frozen set is represented by an arbitrary $r_i \in \Gamma_i^k$, first replace y_i^* by r_i for each frozen coordinate. Proposition 1 preserves $D(y^*)$ through these replacements. The represented tuple then belongs to the current product, and the singleton argument applies unchanged. □

The theorem is pathwise. It assumes that cancellation does not alter the terminal semantics of retained computations. The live-to-shadow check in Section M.3 tests this condition for the evaluated implementation.

C.4 Trajectory-level split-conformal coverage

The statistical layer supplies the coverage event required by the theorem. The fitted categorical probabilities in Equation 4 are held fixed during calibration and testing. For one benchmark source s, let $S_1, \ldots, S_{n_s}$ be the calibration trajectory scores from Equation 5, and let S_{n_s+1} be the corresponding score for a new task group. A group includes both panels, both loads, every registered review epoch, and every identity pending at those epochs.

Proposition 3 (Source-wise trajectory coverage)**.** *Assume the n_s calibration groups and one test group are exchangeable within source s, conditional on the independent fit data. Let*

$$r_s = \lceil (n_s + 1)(1 - \alpha) \rceil \quad \textit{and} \quad q_s = S_{(r_s)}, \tag{11}$$

where $S_{(r_s)}$ is the r_sth ordered calibration score and $r_s \leq n_s$. Then

$$\Pr\left(S_{n_s+1} \leq q_s\right) \geq 1 - \alpha. \tag{12}$$

Consequently, with probability at least $1 - \alpha$, every registered terminal value of the test group lies in its raw set C_i^k.

Proof. Exchangeability makes the rank of the test score among the $n_s + 1$ scores uniform after randomized tie breaking and conservative without it. The probability that the test rank exceeds r_s is at most $(n_s + 1 - r_s)/(n_s + 1) \leq \alpha$. The event $S_{n_s+1} \leq q_s$ implies that each component nonconformity score is at most q_s, because the group score is their maximum. Equation 6 then contains every corresponding terminal value. □

In the initial study, $n_s = 20$, $\alpha = 0.05$, and $r_s = 20$. The finite-sample lower coverage level is therefore $20/21 \approx 95.24\%$ within each source under exchangeability. This statement is conditional on source. The pooled 96-task confidence interval reported in the paper is an empirical summary and does not replace the source-wise conformal premise.

Proposition 4 (Nested-domain coverage)**.** *If $Y_i \in C_i^k$ at every registered epoch for coordinate i, then the update in Equation 7 satisfies $Y_i \in \Gamma_i^k$ at every epoch.*

Proof. The claim holds at the first epoch because $\Gamma_i^1 = C_i^1$. If it holds at epoch $k-1$, then Y_i belongs to both Γ_i^{k-1} and C_i^k. Their intersection is nonempty and contains Y_i, so the first branch of Equation 7 is selected and preserves coverage. Induction completes the proof. □

Combining Propositions 3 and 4 with Theorem 1 yields the decision-risk statement used in the paper.

Proposition 5 (Decision fidelity for the language-model controller)**.** *Under source-wise calibration and test exchangeability, the probability that the controller disagrees with the fixed full-computation decision for a new task group from source s is at most α.*

Proof. The trajectory-level raw coverage event has probability at least $1-\alpha$. Proposition 4 turns that event into coverage by all domains used by the controller. Theorem 1 gives agreement on that event. A disagreement can therefore occur only outside it and has probability at most α. □

C.5 Fixed-priority plurality as a deterministic map

The initial study uses fixed-priority plurality D_{pl}; Section I adds majority with deferral. For valid labels $\mathcal{L}$ and marks $y_i \in \mathcal{L} \cup \{\perp\}$, let $n_a(y) = \sum_i \mathbf{1}\{y_i = a\}$. Abstentions do not contribute to these counts. If at least one valid vote exists, the candidate set is

$$W(y) = \arg\max_{a \in \mathcal{L}} n_a(y). \tag{13}$$

When $|W(y)| > 1$, the chosen label is the tied label supported by the highest-priority identity. If no valid vote exists, the first valid label is returned. Fixed panel priorities make every branch deterministic, so D_{pl} satisfies the premise of Theorem 1.

The guarantee concerns agreement with this complete-panel map. Ground truth does not enter the stopping decision. We report accuracy separately because agreement with the reference and correctness against a benchmark label are different endpoints.

D Counterexamples

D.1 Domain expansion can reverse irrelevance

Nestedness is necessary because expanding another coordinate can create a new relevance witness. Consider two binary coordinates and the map in Table 5. With $G_1 = \{0,1\}$ and $G_2 = \{0\}$, coordinate 1 is irrelevant because the output is always zero. If G_2 later expands to $\{0,1\}$, changing coordinate 1 changes the decision when coordinate 2 equals one. Cancelling coordinate 1 under the smaller domain is therefore unsound if later updates may expand the product.

Table 5: A minimal expansion counterexample. The map is logical conjunction.

y_1	y_2	$D(y_1, y_2)$
0	0	0
0	1	0
1	0	0
1	1	1

The empty-intersection rule in Equation 7 prevents this failure. When a new raw set conflicts with accumulated evidence, the controller keeps the preceding domain. This choice may delay closure, but it never enlarges the domain used to justify an earlier cancellation.

D.2 Terminal irrelevance is not review-time value

Decision relevance and event timing answer different questions. Let $D(y_i, y_j) = y_j$. Coordinate i is irrelevant for every product domain because its terminal value never enters the decision. At time t, suppose the controller records $G_j = \{0,1\}$. Computation i would finish at time $u > t$, computation j at time $v > u$, and no timer review occurs between t and v. If the completion of i triggers a joint

prefix read at u and that read contracts G_j to $\{0\}$, retaining i permits closure at u. Cancelling it removes that review event, so closure is not observed until v.

This example separates decision preservation from resource optimization. Stopping computation i cannot change the terminal action, but it can remove an opportunity to inspect computation j. Delaying that review can also let other calls generate more tokens. The evaluated controller reviews at retained completions, so the resource benefit of freezing relative to global closure must be measured rather than inferred from irrelevance alone.

E Controller and Computational Cost

E.1 Inputs and outputs

The controller receives a fixed decision map, a current prefix or terminal value for each call, a source-specific conformal quantile, and the domains retained from the preceding review. It returns either a final action or a set of calls to continue. Ground-truth labels, future tokens, future completion times, and terminal marks of active calls are not inputs.

At each review epoch, the implementation executes the following sequence.

1. Replace the domain of each newly completed call by its exact parsed terminal mark.
2. Parse every active native prefix and form its raw terminal set with Equation 6.
3. Intersect each active raw set with its previous domain using Equation 7.
4. Enumerate the joint product and compute the decision image in Equation 2.
5. If the image is a singleton, return its unique action and request cancellation of every remaining call.
6. Otherwise, compute Equation 3 for every active coordinate. Freeze and stop each irrelevant call, retain each relevant call, and wait for the next registered review.

The experiment registers natural completion events as reviews. The theory also permits external timer reviews, provided their domains remain nested and the same decision map is used throughout the trajectory.

E.2 Exact finite-domain computation

Let $B_k = \prod_i |\Gamma_i^k|$. Computing the decision attached to each tuple costs B_k evaluations of D. The image follows directly from these evaluations. Coordinate relevance can then be computed by grouping tuples that agree on y_{-i} and testing whether their decisions differ, for total time $O(mB_k)$ beyond the cost of D. Storing the tuple decisions costs $O(B_k)$.

The evaluated panels have five coordinates. The largest task has eight valid labels, so the largest initial terminal alphabet, including abstention, has size nine. The worst unstructured product is thus $9^5 = 59{,}049$ tuples, and contraction rapidly reduces it. Exact enumeration is preferable at this scale because it introduces no relaxation or threshold. Larger panels can use map-specific dynamic programs, decision diagrams, or integer constraints while preserving the same definitions.

E.3 Plurality specialization

For plurality, the image can also be obtained from feasible vote-count vectors and the fixed tie order. The submitted implementation nevertheless uses exact Cartesian enumeration. This keeps the executable object identical to the general definition and makes every relevance decision directly auditable. Cached domain summaries avoid repeating an enumeration when a panel revisits the same label set and domain tuple.

E.4 Stopping and freezing semantics

A global commit and a coordinate freeze are distinct actions. A global commit ends the trajectory because every terminal tuple in the current product maps to the same action. A coordinate freeze ends

one computation because its variation cannot change the image, while the decision may remain open. When a call is frozen, the implementation retains its entire current domain in later image calculations. This choice avoids assigning an artificial terminal mark to a cancelled stream.

F Experimental Protocol

F.1 Evidence partitions

The confirmation data were isolated from method development. Table 6 records the evidence roles. Fit tasks estimate the categorical mapping from a visible prefix mark to a terminal mark. Calibration tasks choose one source-specific trajectory quantile. Confirmation tasks evaluate coverage, decision fidelity, and live resources without updating either object.

Table 6: Disjoint evidence partitions used by each benchmark source.

Partition	Tasks/source	Total tasks	Use
Fit	6	36	Categorical terminal-mark counts
Calibration	20	120	Source-wise trajectory quantile
Development holdout	6	36	Not used to fit or calibrate
Confirmation	16	96	Frozen-method evaluation

Each task remains intact across identities, panels, loads, and review epochs. The shared development cohort has 192 questions, with 32 per source: six fit, twenty calibrate, and six are held outside both roles. We generate these trajectories before freezing the estimator. This grouping prevents the same question from contributing partly to calibration and partly to testing. The 96 confirmation identities have zero overlap with these roles or earlier method-development questions.

F.2 Models and panel construction

Two disjoint panels test whether the controller depends on one model family. Table 7 lists the exact public model repositories and fixed plurality priorities. Priority resolves vote ties only. It does not set launch order, completion order, or a static waiting score.

Table 7: The two disjoint five-model panels. Priority 1 is the highest tie-breaking priority.

Panel	Priority	Short name	Public repository
A	1	SmolLM2-1.7B	`HuggingFaceTB/SmolLM2-1.7B-Instruct`
A	2	OLMo-2-1B	`allenai/OLMo-2-0425-1B-Instruct`
A	3	Granite-3.3-2B	`ibm-granite/granite-3.3-2b-instruct`
A	4	Phi-3.5-mini	`microsoft/Phi-3.5-mini-instruct`
A	5	Mistral-7B-v0.3	`mistralai/Mistral-7B-Instruct-v0.3`
B	1	Qwen3-4B	`Qwen/Qwen3-4B-Instruct-2507`
B	2	Gemma-2-2B	`google/gemma-2-2b-it`
B	3	Qwen2.5-3B	`Qwen/Qwen2.5-3B-Instruct`
B	4	Phi-4-mini	`microsoft/Phi-4-mini-instruct`
B	5	OLMo-2-7B	`allenai/OLMo-2-1124-7B-Instruct`

All five models in a panel are resident during its runs and start for every task. The study uses each repository's chat template, greedy decoding, bfloat16 inference, key-value caching, and a limit of 128 generated tokens. Sampling temperature and nucleus sampling are disabled. The scientific arms do not force a minimum generation length.

F.3 Public benchmarks and confirmation rows

The six sources cover commonsense completion, factuality, broad knowledge, science reasoning, and general commonsense question answering. Table 8 gives the public split and output size. The

source papers define the benchmark tasks and evaluation labels (Clark et al., 2018; Zellers et al., 2019; Hendrycks et al., 2021; Lin et al., 2022; Talmor et al., 2019; Mihaylov et al., 2018).

Table 8: Public confirmation sources. Every source contributes 16 fresh questions.

Source	Public repository	Split	Valid labels
AI2 Reasoning Challenge	`allenai/ai2_arc`	challenge test	4
HellaSwag	`Rowan/hellaswag`	validation	4
Massive Multitask Language Understanding	`cais/mmlu`	all test	4
TruthfulQA MC1	`truthfulqa/truthful_qa`	multiple-choice validation	3 to 8
CommonsenseQA	`tau/commonsense_qa`	validation	5
OpenBookQA	`allenai/openbookqa`	main test	4

For exact row recovery, Table 9 records the selected zero-based row indices within the public revisions stored in the supplementary provenance record. The selection was fixed before model generation and does not use labels or model outcomes.

Table 9: Confirmation row indices within the pinned public dataset files.

Source	Zero-based row indices
AI2 Reasoning Challenge	`5, 57, 61, 100, 109, 257, 306, 421, 429, 489, 536, 799, 859, 873, 892, 1048`
HellaSwag	`41, 118, 145, 2540, 2577, 4086, 4730, 6047, 6189, 6493, 7599, 8421, 8734, 9174, 9361, 9917`
Massive Multitask Language Understanding	`501, 675, 822, 2146, 2164, 3708, 4771, 4957, 5979, 7775, 7923, 8609, 9387, 9638, 12162, 13834`
TruthfulQA MC1	`71, 131, 259, 291, 335, 376, 393, 414, 426, 443, 478, 487, 597, 693, 751, 776`
CommonsenseQA	`48, 114, 121, 497, 526, 625, 653, 658, 827, 840, 873, 906, 1081, 1119, 1160, 1206`
OpenBookQA	`0, 44, 63, 77, 113, 142, 148, 173, 200, 208, 295, 302, 305, 327, 408, 419`

F.4 Prompt and parser

All models receive the same semantic instruction. Models with a system role use separate system and user messages. For a model without that role, the two strings are concatenated in one user message before applying its native chat template.

```
You answer multiple-choice questions.  Reason briefly
if needed.  End with exactly one line in the form FINAL:
<LETTER>, where <LETTER> is one listed option.  Put no text
after that line.
Choose the single best option.
Question:  <question>
Options:
A. <option A>
B. <option B>
...
```

The parser first finds every case-insensitive occurrence of `FINAL: <LETTER>` and accepts the last valid label. If none exists, it accepts a valid label at the start of the first nonempty line when followed by a period, parenthesis, colon, or comma. Every other output maps to $\perp$. The same parser is applied after every newly generated token prefix and after termination. No answer-forcing continuation is added.

F.5 Categorical fit and conformal calibration

The terminal-mark estimator pools identities but conditions on the current parsed mark and the number of valid options. Add-one smoothing gives nonzero mass to every valid label and to abstention. Pooling was frozen during method development because identity-specific cells had much less support and produced wider domains on the development split. Identity remains present in the fixed tie rule and in the coordinate relevance calculation.

Table 10 reports the six frozen trajectory quantiles. Their differences reflect source-specific calibration scores, not tuned risk levels. Every source uses $\alpha = 0.05$ and rank 20 of 20.

Table 10: Source-specific trajectory quantiles fitted before confirmation. Every source uses rank 20 at risk level 0.05.

Source	q_s
ARC-Challenge	0.849879
HellaSwag	0.995968
MMLU	0.849879
TruthfulQA MC1	0.991453
CommonsenseQA	0.971910
OpenBookQA	0.849879

The raw domain always contains labels whose nonconformity score does not exceed the source quantile. If this rule returns no label, the complete terminal alphabet is used. The sequential intersection in Equation 7 then supplies the nested domain used by the controller. Neither the counts nor the quantiles are updated on confirmation tasks.

F.6 Reference decision and physical arms

The full-panel endpoint is abstention-neutral fixed-priority plurality from Section C. Both physical arms launch the same five calls. The full-trace arm records every output and returns D_{pl} after all five calls terminate. The closure arm reads joint prefixes at natural completion events, updates the nested domains, commits on a singleton decision image, and requests cancellation for coordinates outside the relevance set.

Low load executes one task at a time. Every task runs both physical arms, and arm order alternates within each source after a fixed seeded ordering. Each source consequently places eight tasks in each arm order. Batched load uses true per-model batches. Before timing the experimental arms, a capacity probe expands concurrency within an architecture-derived memory envelope and selects feasible concurrency by measured generated-token throughput. The selected capacities were 91 tasks for Panel A and 84 for Panel B. The comparison retains two balanced 48-task clusters per panel, one in each arm order.

The low-load randomization unit is one task. The batched randomization unit is one 48-task cluster. This distinction determines the uncertainty statements in Section G. Batched wall-time effects are reported as order-replicated point effects because two clusters do not support precise inference.

F.7 Endpoints and uncertainty

Decision fidelity is assessed at three levels. Task-level disagreement marks a task as failed if any of its two panels or two loads differs from the same uninterrupted full-trace decision. Raw simultaneous coverage marks a task as failed if any registered pending identity at any review state excludes its terminal mark. Same-trace fidelity uses the uninterrupted trajectory, while live-to-shadow fidelity reconstructs the complete-panel action paired with a cancelled live execution.

Under a common-rate independent Bernoulli model for task failures, the one-sided Clopper–Pearson upper endpoint U for k failures in n task groups solves

$$\sum_{j=0}^{k} \binom{n}{j} U^j (1-U)^{n-j} = 0.05. \tag{14}$$

The pooled binomial bounds summarize the evaluated task mixture; they do not certify conditional coverage for each benchmark source. Low-load resource intervals use 10,000 paired bootstrap draws, resampling task identities within source in each panel. The batched analysis uses each 48-task cluster as one workload unit and reports its two arm-order effects within the same session. Accuracy is reported separately from reference fidelity.

Generated work is the sum of generated tokens over the five calls assigned to a task. Decision latency runs from arm release to the controller decision. Arm wall time runs from arm release until all retained or cancelled workers have terminated. The main resource comparison uses arm wall time because it measures when the physical arm releases its resources. Section G reports both quantities where their distinction matters.

G Complete Confirmation Results

G.1 Decision fidelity and task accuracy

The confirmation study contains 96 task groups, 384 panel-load trajectories, 768 physical arm records, and 3,840 model-worker records. Table 11 gives the complete primary endpoints. Every observed controller decision agrees with its fixed complete-panel reference. The single coverage-failure task is analyzed separately below.

Table 11: Confirmation task-group endpoints and trajectory agreement counts.

Endpoint	Successes	Failures	Failure upper bound
Task decision fidelity	96	0	3.072%
Task raw coverage	95	1	4.846%
Same-trace fidelity	384	0	Task-level bound
Live-to-shadow fidelity	384	0	Task-level bound

The four panel-load trajectories of a task share its question and are correlated. We therefore report their agreement counts and use the task-group bound for uncertainty, without treating the 384 trajectories as independent Bernoulli trials.

Table 12 shows that fidelity equals one in every reported stratum. Accuracy varies by panel and source, as expected for heterogeneous small models, but it is exactly equal for closure and complete-panel plurality within every row.

Table 12: Confirmation behavior by panel, load, and source. Closure and full-reference accuracy agree in every row.

Stratum	Traces	Fidelity	Accuracy	Closure index
Panel A, all loads	192	100.0%	64.06%	2.76
Panel B, all loads	192	100.0%	76.56%	3.22
Panel A, low load	96	100.0%	63.54%	2.55
Panel A, batched	96	100.0%	64.58%	2.96
Panel B, low load	96	100.0%	78.12%	3.22
Panel B, batched	96	100.0%	75.00%	3.23
ARC-Challenge	64	100.0%	84.38%	2.22
HellaSwag	64	100.0%	68.75%	4.19
MMLU	64	100.0%	68.75%	2.83
TruthfulQA MC1	64	100.0%	48.44%	4.08
CommonsenseQA	64	100.0%	78.12%	2.16
OpenBookQA	64	100.0%	73.44%	2.47

G.2 Low-load live resource effects

The low-load comparison provides the strongest resource inference because each panel has 96 paired task units. Table 13 reports arm means, paired differences, intervals, and relative reductions. The generated-token endpoint and the physical arm wall-time endpoint both favor closure in both panels.

Table 13: Low-load paired effects. Differences are closure minus full, with source-stratified 95% paired bootstrap intervals over 96 tasks per panel.

Panel	Endpoint	Full	Closure	Difference (95% interval)	Saving
A	Tokens	148.97	75.53	-73.44 (-83.59, -63.26)	49.3%
A	Wall time (s)	1.59	0.84	-0.75 (-0.86, -0.65)	47.4%
B	Tokens	179.65	113.11	-66.53 (-78.75, -54.83)	37.0%
B	Wall time (s)	2.30	1.50	-0.80 (-0.96, -0.66)	34.9%

The effect appears across all six sources, although the amount depends on how early their panels become determined. Table 14 gives descriptive source means under low load. Each cell is a relative reduction from the paired full arm. HellaSwag and TruthfulQA close later and therefore offer less remaining work to remove, especially in Panel A for HellaSwag and in Panel B for TruthfulQA.

Table 14: Low-load reductions by source, with 16 paired tasks per panel.

Source	A tokens	A time	B tokens	B time
ARC-Challenge	66.9%	63.5%	51.4%	48.7%
HellaSwag	11.4%	11.2%	31.8%	32.9%
MMLU	63.4%	59.0%	25.8%	23.0%
TruthfulQA MC1	33.4%	31.6%	13.5%	12.0%
CommonsenseQA	68.3%	67.2%	53.6%	47.3%
OpenBookQA	70.2%	67.8%	52.9%	50.2%

Both arm orders retain the low-load effect. When closure runs first, token reductions are 51.4% and 41.8% for Panels A and B, and wall-time reductions are 49.2% and 38.3%. When full execution runs first, token reductions are 47.4% and 32.6%, and wall-time reductions are 45.8% and 31.6%. The effect therefore has the same direction in both order halves.

G.3 Batched-load stress test

The batched condition tests whether savings remain observable when many queries share each model. Table 15 reports two cluster contrasts per panel. Generated work falls by nearly the same fraction as under low load. Wall-time reductions are smaller because cancellation competes with batching, queue completion, and fixed arm overhead.

Table 15: Initial batched-load effects. Ranges span the two workload reductions and are not confidence intervals.

Panel	Endpoint	Pooled saving	Workload range
A	Tokens	50.0%	46.2% to 53.8%
A	Wall time	9.3%	8.2% to 10.2%
B	Tokens	36.3%	34.9% to 37.8%
B	Wall time	3.3%	1.8% to 4.9%

All eight workload-level reductions are positive. With only two units per panel, however, a two-sided exact sign test is necessarily $p = 0.5$. The result establishes directional replication across arm order, not a precise batched effect size.

G.4 Closure and observed completions

Figure 2 separates the reference-trace closure index from the number of completions physically observed by the live controller. Decision closure occurs before the fifth completion on 291 of 384 reference traces. Selective freezing removes future completion events, so the live controller observes fewer than five completions on 331 trajectories.

The difference between the two rows is expected. Closure index is defined on the uninterrupted reference trace. The live controller may stop an irrelevant call before its natural completion and

Table 16: Completion counts over 384 confirmation trajectories.

Completions	1	2	3	4	5
Reference closure	121	31	56	83	93
Live observed	124	78	65	64	53

therefore never observe that event. This distinction also explains why the mean reference closure index, 2.99, is larger than the mean number of observed live completions, 2.59.

G.5 Model-level heterogeneity

The panels contain large differences in correctness, abstention, output length, and completion time. Table 17 uses one low-load full-trace execution per task and model, giving 96 records for every row. These values describe the panel members; they are never used as static cancellation scores.

Table 17: Low-load full-trace model behavior over 96 tasks. Finish time starts at simultaneous arm release.

Panel	Model	Accuracy	Abstention	Tokens	Finish (s)
A	Granite-3.3-2B	53.12%	5.21%	32.44	1.07
A	Mistral-7B-v0.3	63.54%	0.00%	25.66	0.94
A	OLMo-2-1B	15.62%	12.50%	12.96	0.53
A	Phi-3.5-mini	70.83%	1.04%	64.79	1.45
A	SmolLM2-1.7B	48.96%	0.00%	13.12	0.55
B	Gemma-2-2B	57.29%	1.04%	10.20	0.68
B	OLMo-2-7B	58.33%	10.42%	11.65	0.77
B	Phi-4-mini	76.04%	1.04%	17.92	0.84
B	Qwen2.5-3B	76.04%	2.08%	44.98	1.43
B	Qwen3-4B	56.25%	36.46%	94.91	2.29

The table illustrates why relevance cannot be reduced to speed or average accuracy. Phi-3.5-mini is the slowest member of Panel A but is also its most accurate individual member. Qwen3-4B generates the longest outputs and has a high abstention rate under the strict parser. Depending on the current joint domains and votes, either model may be essential, irrelevant, or already dominated by a closed decision.

G.6 Observed coverage failure

The only raw simultaneous-coverage failure occurs on OpenBookQA row 327. In Panel A, Mistral's visible mark is D at four low-load and three batched-load reviews, while its terminal mark is A. The raw domain is $\{D\}$ and misses A at all seven rows. These rows form one failure under the task-group endpoint.

This failure does not change the aggregate decision. Both the closure controller and complete-panel plurality return C in every affected trajectory. The observation is consistent with the one-way implication in Proposition 5: coverage guarantees decision fidelity, but a missed coordinate can remain irrelevant to the final map.

Per-source test sets are too small to establish a 5% empirical upper bound. The registered pooled task result, one failure among 96 groups, has the 4.846% upper endpoint reported in the main text. Source-wise conformal validity follows from the calibration rank argument; the small test subsets provide descriptive checks.

G.7 Physical cancellation

The live controller issued 924 cancellation requests, of which 923 were observed by workers. A cancellation request can coincide with natural termination, so request and observation are distinct counters. Every worker terminated without error, and the maximum generation after observing a

Table 18: Task-group raw coverage by source. One-sided 95% binomial bounds are descriptive summaries over 16 questions.

Source	Covered	Failed	Failure upper bound
ARC-Challenge	16	0	17.07%
HellaSwag	16	0	17.07%
MMLU	16	0	17.07%
TruthfulQA MC1	16	0	17.07%
CommonsenseQA	16	0	17.07%
OpenBookQA	15	1	26.40%

cancellation flag was one token step. Table 19 separates precommit freezes from requests made after global closure.

Table 19: Cancellation by identity in confirmation. Precommit and global-closure requests are distinct actions.

Panel	Model	Precommit	At closure	Requested
A	Granite-3.3-2B	44	82	126
A	Mistral-7B-v0.3	8	100	108
A	OLMo-2-1B	16	56	72
A	Phi-3.5-mini	33	128	161
A	SmolLM2-1.7B	5	13	18
B	Gemma-2-2B	44	42	86
B	OLMo-2-7B	22	50	72
B	Phi-4-mini	38	42	80
B	Qwen2.5-3B	0	76	76
B	Qwen3-4B	0	125	125

These totals are state-dependent outcomes, not a ranking of model value. A model with no precommit freeze can still be cancelled after closure, and a frequently frozen model can remain relevant on a different task.

H Mechanism Analysis on Fixed Traces

H.1 Four matched policies

The mechanism analysis replays four policies on the same 384 uninterrupted confirmation traces. Full panel waits for all five terminal marks. Logical closure gives every pending call its unrestricted terminal alphabet and commits when completed votes alone force the decision. Prefix-domain closure uses the nested conformal domains but never freezes an active coordinate before global closure. Selective freezing adds Equation 3 to prefix-domain closure.

Because the traces are fixed, this analysis identifies logical and token-accounting mechanisms without introducing scheduling variation. It does not replace the randomized physical-arm comparison in Section G.

Table 20: Matched policy decomposition over 384 uninterrupted traces. All decisions preserve their references; timing is measured on fixed traces.

Policy	Closure index	Tokens	Token saving	Trace-time saving
Full panel	5.00	164.65	0.0%	0.0%
Logical closure	4.01	118.05	28.3%	32.0%
Prefix domains	2.92	95.50	42.0%	41.8%
Selective freezing	2.99	92.27	44.0%	41.0%

The table isolates the central mechanism. Completed votes alone remove 28.3% of generated work. Native prefix domains add 13.7 percentage points. Coordinate freezing adds another 2.0 points. The

slight increase in closure index and trace time from prefix-domain closure to selective freezing is consistent with the review-event effect in Section D.

H.2 Uncertainty in the additional token savings

The incremental effects are stable under task-level resampling of the recorded traces. This analysis uses 10,000 paired bootstrap draws. Within each source, it samples 16 task identities with replacement and retains all four panel-load executions of each selected task. For each draw, it divides the difference in total generated tokens between adjacent policies by the total full-panel tokens in that same draw. This estimates percentage-point gains in reduction relative to full-panel execution while preserving dependence between policies and executions of the same task.

Table 21: Additional token reduction in percentage points. Intervals resample task identities within source.

Contrast	Gain	95% lower	95% upper
Prefix over completed votes	13.69	11.11	16.44
Freezing over prefix-global	1.96	1.15	2.91

Prefix domains account for 87.5% of the extra saving beyond completion-only logical closure. The two interval estimates in Table 21 exclude zero. They quantify task variation in token accounting conditional on the observed executions.

Table 22: Source-level incremental token reduction in percentage points.

Source	Prefix increment	Freezing increment
ARC-Challenge	22.62	2.29
HellaSwag	0.00	0.00
MMLU	19.90	4.47
TruthfulQA MC1	5.82	-0.44
CommonsenseQA	13.38	3.69
OpenBookQA	24.76	2.38

The source breakdown distinguishes useful prefix information from completion-only stopping. Five sources gain from calibrated prefixes, while HellaSwag retains the complete five-symbol terminal alphabet in every pending state. Its three early-stopping policies consequently agree in generated work. This provides an observed limiting case of the method: when prefixes exclude no terminal value, the domain controller reduces to logical closure from completed votes.

H.3 Panel and load decomposition

The same ordering appears across panels and loads, with a different contribution from prefix information in each panel. Table 23 reports token reductions from the fixed traces. Selective freezing improves generated work in all four strata, even when its closure index is unchanged or slightly later.

Table 23: Generated-token reductions on matched traces by panel and load.

Stratum	Completed votes	Prefix domains	Selective freezing
Panel A, low	23.8%	49.6%	51.4%
Panel A, batched	23.9%	48.9%	51.1%
Panel B, low	32.5%	36.5%	38.4%
Panel B, batched	31.4%	35.4%	37.4%

Panel A benefits more from prefix domains, while freezing adds roughly two percentage points in either panel. These differences follow from the observed vote, domain, and completion configurations. They do not imply that either panel has an intrinsically better controller.

H.4 Domain informativeness

Raw prediction sets are informative when a visible mark is stable in the fit traces and wider when the prefix is unparsed. Table 24 summarizes all registered pending rows. A singleton raw set appears often for AI2 Reasoning Challenge, Massive Multitask Language Understanding, CommonsenseQA, and OpenBookQA. HellaSwag and TruthfulQA produce wider sets, which is consistent with their later closure in Table 12.

Table 24: Raw terminal domains across pending states. Task-group coverage is reported separately.

Source	Pending rows	Mean size	Singleton fraction
ARC-Challenge	640	2.76	56.1%
HellaSwag	640	5.00	0.0%
MMLU	640	3.12	46.9%
TruthfulQA MC1	640	5.56	7.8%
CommonsenseQA	640	3.95	41.1%
OpenBookQA	640	2.88	53.1%

The set construction remains useful even when one coordinate has a wide domain. Decision closure depends on the image of the entire product. Completed exact marks and narrower domains from other calls can make a wide coordinate irrelevant or make all its possible values lead to the same action.

H.5 Selective freezing and closure timing

Prefix-domain global closure and selective freezing agree at the same completion index on 362 of 384 trajectories. Selective freezing closes later on 22 and earlier on none. It nevertheless removes 3.23 additional generated tokens per trajectory and avoids 4,259 token steps before global commitment. The net reduction is smaller than the precommit saving because freezing can postpone later reviews. In these initial traces, freezing saves additional work, while prefix-informed global closure supplies the principal reduction in decision time.

A scheduler can preserve selected event sources after their terminal value becomes irrelevant, or add timer reviews that do not depend on call completion. Either choice changes the review protocol while retaining the same decision certificate. Its latency effect requires a separate live intervention.

I Independent Extension with Frozen Domains

I.1 Questions and statistical units

The extension tests whether the initial result persists on new questions and under a second decision rule. It registers 48 additional questions from each of the six sources, for 288 distinct questions. All selected identities are disjoint from earlier fit, calibration, development, and confirmation data. Model repositories, prompts, parser, 128-token limit, categorical counts, source quantiles, and panel priorities remain fixed. Thus the extension tests the existing controller without refitting to the new outcomes.

The workload separates question coverage from repeated resource measurements. Every new question runs in both panels at high load. Eight questions per source, selected before execution, also run at low load in each panel. The resulting low-load comparison has 48 paired questions per panel. Capacity calibration selects batches of 62 for Panel A and 54 for Panel B. Each panel executes eight high-load workload pairs, with four pairs in each arm order. The schedule first covers all questions and fills later batch positions with repeated identities. Both arms of every pair use the same microbatch sizes and runtime session.

Each analysis uses the unit appropriate to its question. Fidelity counts a question once if any evaluated execution disagrees with its full reference. The primary decision-rule replay uses the first high-load trace of each question in each panel, giving 576 equally weighted trajectories. This avoids giving extra weight to the 48-question low-load subset or to repeated batch entries. Live resource effects use the ratio of total closure-arm cost to total full-arm cost within each panel and load. Low-load

intervals resample paired question identities within source over 10,000 bootstrap draws. High-load ranges describe the eight observed workload pairs in one session; repeated questions do not increase the number of independent tasks.

The decision-rule tables use the 576-trace summary consistently. The result archive also retains the low-load executions and all repeated high-load traces, so resource totals include every scheduled workload without changing the question weights in the rule comparison.

Table 25: Extension design. Repeated executions do not add independent questions.

Object	Count and role
New questions	288, with 48 per source
Primary rule traces	576, one per question and panel
Low-load pairs	48 per panel
High-load workload pairs	Eight per panel, four per arm order
Reference executions	1,024, including repetitions
Worker records	10,240, across both arms

I.2 Reference fidelity and the observed disagreement

The frozen controller preserves the reference on 287 of 288 new questions. One Panel A execution changes the plurality decision; all other observed executions agree. Live and separately reconstructed shadow actions match on all 1,024 reference trajectories, including the disagreement. The empirical task-level error rate is 0.35%, with a descriptive one-sided 95% Clopper–Pearson upper limit of 1.64%. This pooled sample summary is separate from source-wise conformal coverage. In particular, the new batching schedule changes which prefixes are observed and tests the transfer of the frozen quantiles empirically.

Table 26: Extension fidelity counts. Task counts merge repeated executions.

Endpoint	Observation
Tasks with disagreement	1/288
Tasks with raw coverage miss	4/288
Primary reference matches	575/576
All reference matches	1,023/1,024
Live-to-shadow matches	1,024/1,024

The disagreement identifies a failure of domain coverage at the stopping state. On ARC-Challenge test row 884, the domains for OLMo-2, SmolLM2, Granite, and Phi-3.5 are $\{B\}$, $\{B\}$, $\{D\}$, and $\{C\}$, while Mistral's domain remains unrestricted. B wins every compatible tuple, including a possible tie with C or D, because OLMo-2 has first priority. The controller therefore returns B. In the full execution, Granite and Mistral finish with C, producing three C votes. Full plurality returns C, the correct answer.

This case distinguishes statistical coverage from execution correctness. The controller applies the declared decision map correctly to its current product domain, and the live action matches the shadow action. However, Granite's terminal C lies outside its domain, so the joint coverage premise of Theorem 1 is false on this trajectory. Three additional questions contain raw coverage misses without changing the aggregate decision. No quantile, threshold, or parser was changed after observing these outcomes.

I.3 Live resource measurements

The extension reproduces substantial low-load reductions in both panels. Table 27 reports the complete paired comparison. Panel A saves 56.8% of generated tokens and 53.6% of wall time; Panel B saves 41.3% and 39.0%, respectively. All four source-stratified bootstrap intervals exclude zero. These measurements extend the initial low-load result to a new balanced question subset with the same controller.

Table 27: Extension low-load effects with source-stratified 95% paired bootstrap intervals, over 48 tasks per panel.

Panel	Endpoint	Saving	95% lower	95% upper
A	Tokens	56.8%	51.6%	61.9%
A	Wall time	53.6%	47.4%	59.4%
B	Tokens	41.3%	33.3%	49.0%
B	Wall time	39.0%	30.8%	46.8%

High-load execution retains a larger token benefit than wall-time benefit. Table 28 reports the pooled reduction, the observed workload range, and how many pairs favor closure. All eight token comparisons favor closure in each panel. Wall time improves in all eight Panel A pairs and seven of eight Panel B pairs. The slower Panel B workload remains in every reported total. This pattern agrees with the initial study: eliminating generated work does not remove the shared batch's longest retained computation.

Table 28: Extension high-load effects. Ranges describe eight observed workload pairs, not confidence intervals.

Panel	Endpoint	Pooled saving	Workload range	Favorable pairs
A	Tokens	49.1%	41.5% to 54.0%	8/8
A	Wall time	8.8%	0.0% to 24.6%	8/8
B	Tokens	39.5%	35.6% to 47.5%	8/8
B	Wall time	1.0%	-2.7% to 2.7%	7/8

These repetitions establish conditional resource savings for the recorded workloads. They share a runtime session and reuse some question identities. Consequently, they measure variation across scheduled batches within that session, rather than variation across independent deployments. The larger set of new questions strengthens the fidelity evaluation; the additional workload pairs strengthen the execution check without changing its inferential unit.

I.4 Strict majority with deferral

The second decision map tests whether the same terminal domains can support an action beyond answer selection. Let $\mathcal{L}$ be the valid label set and let $v_\ell(y) = \sum_{i=1}^{5} \mathbf{1}\{y_i = \ell\}$. Define

$$D_{\mathrm{maj}}(y) = \begin{cases} \ell, & v_\ell(y) \geq 3, \\ \mathrm{REVIEW}, & \max_{\ell \in \mathcal{L}} v_\ell(y) < 3. \end{cases} \tag{15}$$

At most one label can have three votes. Abstention casts no vote and never reduces the five-model denominator. The action REVIEW denotes deferral for a later review process. We measure agreement with this fixed action and the computation needed to determine it; no human review outcome or cost is included.

The decision-image test applies to this map without changing the prediction sets. Completion-only closure uses unrestricted pending domains, while prefix closure uses the frozen conformal domains and their nested intersections. Both stop when the image under D_{maj} is a singleton. This replay tests global closure at the recorded review epochs. Its token savings are counterfactual trace quantities; the physical-arm resource measurements above use the original plurality controller.

Prefix domains add 16.5 percentage points of token reduction for majority with deferral. All 576 actions agree with the full rule, including the 119 deferrals, or 20.7% of trajectories. Full majority and prefix closure each return a correct label on 366 trajectories. These counts exclude deferred actions from correct labels and make no assumption about their later resolution.

The source pattern is consistent with the initial mechanism study. Five sources gain from prefix domains, while HellaSwag receives the same saving from completed votes alone. The domains therefore retain their empirical role when the decision map changes: useful prefix information reduces work, and uninformative prefixes leave the controller to rely on completed outputs.

Table 29: Two fixed maps on 576 equally weighted extension traces. Each policy is compared with its own full rule.

Rule	Evidence	Token saving	Reference matches
Plurality	Completed votes	28.2%	576/576
Plurality	Prefix domains	43.7%	575/576
Majority with deferral	Completed votes	26.5%	576/576
Majority with deferral	Prefix domains	43.1%	576/576

Table 30: Majority-with-deferral replay by source. Each row has 96 panel traces.

Source	Vote saving	Prefix saving	Full deferral	Errors
ARC-Challenge	29.0%	58.0%	16.7%	0
HellaSwag	17.8%	17.8%	29.2%	0
MMLU	21.6%	42.8%	31.2%	0
TruthfulQA MC1	27.0%	36.2%	18.8%	0
CommonsenseQA	34.3%	53.4%	15.6%	0
OpenBookQA	35.6%	59.2%	12.5%	0

I.5 Native-margin diagnostics

Two registered diagnostics test a margin criterion using the same native marks. For pending call i with mark m_i, the estimated switch probability is $q_i = 1 - \widehat{p}(m_i \mid m_i, L)$ from Equation 4; completed calls have $q_i = 0$. Let c be the current fixed-priority plurality winner, v_c its vote count, and u the number of pending calls. A diagnostic stops only when

$$v_c > \sum_i q_i\big(1 + \mathbf{1}\{m_i = c\}\big) + b, \tag{16}$$

$$v_c - v_\ell > \sum_i q_i\big(1 + \mathbf{1}\{m_i = c\} - 2\mathbf{1}\{m_i = \ell\}\big) + b \quad \text{for every voted label } \ell \neq c. \tag{17}$$

The first inequality also tests a challenger with no current votes. The plug-in variant uses $b = 0$; the corrected variant uses $b = \sqrt{2u \log(100L)}$, with the factor 100 corresponding to the registered 0.05 budget over five reviews. Both return the full reference if all calls complete. This correction is evaluated as a diagnostic; the experiment does not establish the independent-switch assumptions needed to interpret it as a risk bound.

Table 31: Native-margin diagnostics on 576 extension traces. These local constructions are not official end-to-end stopping systems.

Diagnostic	Token saving	Reference matches	Failed tasks
Plug-in	58.6%	566/576	10/288
Corrected	1.4%	576/576	0/288

The two diagnostics illustrate the cost of their respective stopping conditions. The plug-in criterion saves more tokens and changes ten question-level reference decisions. The correction preserves all observed decisions while removing little work. The domain controller occupies an intermediate observed operating point (Table 29); these results alone do not identify an optimal risk-cost tradeoff.

I.6 Scope of the extension

The extension strengthens question coverage, live replication, and the range of tested decisions. It adds 288 questions without refitting and instantiates the fixed-map theorem with an explicit deferral action. Together, the two studies cover 384 distinct questions, ten models, and six public multiple-choice benchmarks. They do not evaluate the entire benchmark test sets or new modalities.

Comparison with recent parallel stopping systems remains a separate experiment. Parallel-Probe and MARS use their own answer probes, review schedules, and calibration procedures (Zheng et al., 2026;

Chen et al., 2026). The extension records native-margin adaptations as diagnostics, whose estimates use the frozen categorical fit. These are not executions of the official systems, so they cannot establish a performance ranking against them. A matched end-to-end comparison would require accounting for those systems' probe and calibration costs under a common generation budget.

J Complete OpenBookQA Test Evaluation

The full-test evaluation extends the evidence beyond benchmark subsets. Both five-model panels run on all 500 questions in the OpenBookQA main test split (Mihaylov et al., 2018). The controller preserves both panel decisions on 499 questions. Generated tokens fall by 71.0% in Panel A and 49.0% in Panel B. Validation-only fitting and calibration provide a separate domain model for this evaluation.

J.1 Data separation and validation-only calibration

The official validation split supplies all fitted parameters for this evaluation. We shuffle its 500 zero-based row indices with seed 20260909, assign the first 32 to categorical fitting, and assign the next 64 to calibration. Both panels generate a full trace for each selected question. Fit and calibration identities are disjoint, and the selected validation content does not duplicate test content. No test output updates either the categorical counts or the quantile. The two stages require 192 full-trace panel executions before testing.

Calibration retains the task-level grouping used in the main method. Each of the 64 calibration scores is the maximum nonconformity over both panels and all observed pending states at completion reviews, using counts from the 32 fit questions. With $\alpha = 0.05$, the split-conformal rank is $\lceil 65(1 - 0.05) \rceil = 62$. The resulting quantile is 0.907407. Table 32 gives every fitted count and raw domain. This construction uses the same add-one estimate and nested update as Equation 4 and Equation 7. The risk level and stopping rule are unchanged.

Table 32: Validation-fitted terminal counts and four-option raw domains. Columns are terminal marks; each row is a visible prefix mark.

Prefix	A	B	C	D	$\perp$	Raw domain
A	72	4	0	0	0	$\{A\}$
$\perp$	71	62	65	92	29	$\{A, B, C, D, \perp\}$
B	0	75	0	0	0	$\{B\}$
C	0	0	81	0	0	$\{C\}$
D	3	0	0	86	0	$\{D\}$

The complete test split contains two distinct history groups. Earlier fitting, calibration, or evaluation used 96 of its questions; the other 404 were previously unused. We register these groups before generation and report both alongside the full split. Validation-only fitting separates this study's parameter estimation from its test outputs, while the history groups retain the distinction between prior method development and additional question coverage. The source-wise exchangeability condition remains the one stated in Equation 8 (Vovk et al., 2005; Angelopoulos & Bates, 2023).

J.2 Paired execution on the full split

Every test question receives one full-trace arm and one actual-cancellation arm in each panel. The model identities, fixed priority orders, prompts, parser, greedy decoding, and 128-token limit match the initial study. Throughput and memory probes use validation prompts; test prompt lengths enter the memory estimate. The resulting maximum batches contain 196 questions in Panel A and 171 in Panel B. The final batches contain 108 and 158 questions, respectively, without repeated padding questions.

The six workload pairs use randomized arm order within each panel, with order counts differing by one. Both arms of each pair run in the same session with identical question groups and batch sizes. Each panel contributes three pairs. Each question occurs once per panel, yielding 1,000 full traces

and 1,000 cancellation executions. Completed validation and test work is checkpointed separately, with calibration fixed before test generation. The complete package contains 2,192 task-arm records and 10,960 model-task records.

J.3 Reference fidelity and resource reductions

The full-test result preserves reference decisions on 499 of 500 questions. Panel A agrees on all 500 executions, and Panel B agrees on 499. Every actual-cancellation decision matches the separately reconstructed closure decision on its paired reference trace. Five questions have at least one raw coverage miss. Table 33 reports these quantities for the complete split and both history groups, with a question counted as preserved only when both panels agree with their references.

Table 33: Full-test fidelity and raw coverage by prior-use group. Both panel references must be preserved for a task to match.

Group	Questions	Reference matches	Raw misses	Live/shadow
Complete test	500	499/500	5	1,000/1,000
Previously unused	404	403/404	3	808/808
Previously used	96	96/96	2	192/192

Both panels reduce generation while retaining their full-panel accuracy. Panel A generates 19,262 tokens under cancellation versus 66,401 under full execution; Panel B generates 42,530 versus 83,388. Reference and controller accuracy are 73.6% for Panel A and 84.8% for Panel B. The single changed decision substitutes one incorrect answer for another, so reference disagreement and accuracy loss differ on that question. Table 34 reports the paired resource and accuracy results.

Table 34: Complete-test accuracy and live resources. Token counts are means per question; correct-answer counts cover the full test. Time saving uses paired workload totals.

Panel	Full correct	Stop correct	Full tokens	Stop tokens	Token saving	Time saving
A	368	368	132.80	38.52	71.0%	25.0%
B	424	424	166.78	85.06	49.0%	6.2%

The previously unused questions show the same benefit. On these 404 questions, Panel A reduces generated tokens by 71.4% and Panel B by 49.4%, with unchanged accuracy in both panels (Table 35). The one reference disagreement occurs in Panel B. Thus the full-split reduction is also present outside the questions used in earlier work.

Table 35: Test history groups. Token saving is relative to each group's full execution; mixed-batch wall time is not attributed to subgroups.

Group	Panel	Questions	Reference matches	Full / stop correct	Token saving
Unused	A	404	404/404	296/296	71.4%
Used	A	96	96/96	72/72	69.5%
Unused	B	404	403/404	339/339	49.4%
Used	B	96	96/96	85/85	47.1%

Every physical workload pair favors cancellation in wall time. Summing paired workload durations gives 66.01 versus 49.53 seconds for Panel A and 61.56 versus 57.77 seconds for Panel B, corresponding to reductions of 25.0% and 6.2%. Table 36 includes all six pairs and the partial final batches. These are complete-workload measurements in one session. They describe this execution condition, rather than independent replications across deployment sessions. Because each batch mixes history groups, its wall time cannot be attributed separately to the 404-question subgroup.

J.4 Global closure and a second decision rule

Prefix domains retain their advantage over completed votes on the full test split. On the same 1,000 uninterrupted traces, completed-vote closure saves 35.6% of tokens for plurality, while prefix-

Table 36: Every paired test workload. Order indicates the first arm. Wall times cover whole workloads within one runtime session.

Panel	Tasks	First arm	Full (s)	Stop (s)	Token saving	Time saving
B	171	Full	21.40	20.13	48.8%	6.0%
B	171	Stop	19.78	19.38	46.8%	2.0%
B	158	Full	20.38	18.26	51.4%	10.4%
A	196	Full	25.79	18.34	69.5%	28.9%
A	196	Full	24.58	21.06	71.5%	14.3%
A	108	Stop	15.64	10.13	72.7%	35.2%

informed global closure saves 57.5%. Both preserve every reference decision. This 22.0-point difference measures the contribution of observing unfinished prefixes without selective freezing. It is a fixed-trace token comparison; Table 34 measures the complete controller's physical execution.

The same domains also support strict majority with deferral. Completed-vote closure saves 33.2% of tokens and preserves all reference actions. Prefix-informed closure saves 56.5% and preserves 998 of 1,000 actions, with 145 deferrals compared with the full rule's 147. Correct labels increase from 730 to 731, counting deferrals as incorrect. Table 37 reports these policies together with the two previously specified native-margin diagnostics. The plug-in diagnostic saves 58.5% with 14 changed reference decisions, whereas its corrected form saves 1.7% with none. Their tradeoff concerns the local adaptations, not the official MARS or Parallel-Probe systems.

Table 37: Complete-test fixed-trace comparison. Each stopping policy is compared with its own full-computation decision.

Rule	Policy	Token saving	Reference matches	Deferral
Majority	Completed votes	33.2%	1,000/1,000	14.7%
Majority	Prefix domains	56.5%	998/1,000	14.5%
Plurality	Completed votes	35.6%	1,000/1,000	0.0%
Plurality	Margin corrected	1.7%	1,000/1,000	0.0%
Plurality	Margin plug-in	58.5%	986/1,000	0.0%
Plurality	Prefix domains	57.5%	1,000/1,000	0.0%

Both majority-rule changes replace deferral with a label. On row 392 in Panel B, prefix-global closure returns the correct label B instead of REVIEW. On row 155 in Panel A, it returns D instead of REVIEW, while the correct label is B. Both stopping paths exclude a terminal mark from a visited domain. These two cases explain the lower deferral count and one additional correct label without treating correctness as reference agreement.

J.5 Coverage and the changed reference decision

The single decision change occurs after a raw domain excludes the model's eventual output. On test row 293, Panel B finishes with Qwen3 abstaining, Gemma answering A, Qwen2.5 answering D, Phi-4 answering C, and OLMo-2 answering D. Full plurality therefore selects D. At the first review, however, Qwen2.5's prefix is C and its raw domain is $\{C\}$. That domain excludes the eventual D. Table 38 lists every raw coverage miss, including the four questions that retain the reference decision.

Table 38: All raw coverage-miss identities. Repeated pending states are counted in the last column; task failures merge them.

Test row	Panel	Model	Prefix	Terminal	Group	States
155	A	Mistral-7B-v0.3	D	A	Unused	2
293	B	Qwen2.5-3B	C	D	Unused	3
304	B	Gemma-2-2B	B	C	Used	1
327	A	Mistral-7B-v0.3	D	A	Used	4
392	B	Qwen2.5-3B	B	D	Unused	3

Selective freezing retains the incorrect singleton after the call is cancelled. At the first review, Phi-4 has finished with C. Gemma, Qwen2.5, and OLMo-2 have singleton domains $\{A\}$, $\{C\}$, and $\{D\}$. Their coordinates are invariant within those domains, so the controller freezes them. Qwen3 later abstains, leaving C as the selected decision. The ground-truth label is A, so both the reference D and controller C are incorrect. Global prefix closure without freezing waits for Qwen2.5's terminal D and recovers the reference decision on this trace.

This case separates conditional validity from observed domain coverage. The freezing calculation is exact for its current domains, but the true terminal tuple has already left their product. The premise of Theorem 1 therefore fails. The result reinforces the role of calibrated terminal uncertainty: smaller domains can remove more work, and their coverage determines whether that action preserves the reference. No quantile or parser was changed after observing the test outcomes.

K Complete ARC-Challenge Test Evaluation

The second full-test evaluation reproduces resource savings on a larger science-question benchmark. Both panels run on all 1,172 official ARC-Challenge test questions (Clark et al., 2018). They preserve both reference decisions on 1,170 questions, reduce generated tokens by 67.8% and 50.9%, and reduce paired workload time by 10.8% and 4.3%. This study uses its own validation-only fit and calibration, with the same method and risk level as the OpenBookQA evaluation.

K.1 Complete test coverage and independent validation

The official validation split supplies all estimated parameters. We shuffle its 299 zero-based indices with seed 20260910, use the first 32 questions for categorical fitting, and use the next 64 for calibration. The selected roles are disjoint and have no exact question-and-option duplicates across roles. Both panels generate a complete trace for each validation question. Test outcomes enter neither the counts nor the calibration scores.

Calibration uses one maximum score per question across both panels and all observed pending completion-review states. With $\alpha = 0.05$, the rank is $\lceil 65(1 - 0.05) \rceil = 62$, giving quantile 0.827243. Table 39 reports all fitted counts and their resulting four-option raw domains. The source-wise exchangeability condition and nested update are those of Equation 8 and Equation 7 (Vovk et al., 2005; Angelopoulos & Bates, 2023).

Table 39: Validation-fitted terminal counts and four-option raw domains. Columns are terminal marks; each row is a visible prefix mark.

Prefix	A	B	C	D	$\perp$	Raw domain
A	64	0	0	2	0	$\{A\}$
$\perp$	57	63	70	55	51	$\{A, B, C, D, \perp\}$
B	0	90	0	0	0	$\{B\}$
C	0	0	92	0	0	$\{C\}$
D	0	3	0	93	0	$\{D\}$

All official test questions are retained, including uncommon option counts. The split contains four three-option questions, 1,165 four-option questions, and three five-option questions. Native answer labels are mapped by option position to consecutive letters without changing option text or order. The selected fit and calibration questions happen to have four options. For unseen option counts, Equation 4 assigns uniform mass to the valid labels and abstention; the raw-set rule returns the full terminal domain. Those questions therefore use completed-vote closure until enough exact outcomes arrive. Table 40 reports each option-count group.

The history groups distinguish full-split coverage from new-question evidence. Earlier development or evaluation used 256 test questions, leaving 916 previously unused. These groups were specified before generation, and both receive the new validation-fitted controller. Earlier cancellation records do not replace the new paired executions because their fitted domains differ. Table 41 reports the complete split and both groups.

Table 40: Complete ARC-Challenge test coverage by option count. Both panel references must match for task fidelity.

Options	Questions	Reference matches	A token saving	B token saving
3	4	4/4	38.8%	44.1%
4	1165	1,163/1,165	67.9%	50.9%
5	3	3/3	33.5%	42.2%

Table 41: Full-test fidelity and raw coverage by prior-use group. Both panel references must be preserved for a task to match.

Group	Questions	Reference matches	Raw misses	Live/shadow
Complete test	1172	1,170/1,172	16	2,344/2,344
Previously unused	916	915/916	12	1,832/1,832
Previously used	256	255/256	4	512/512

K.2 Matched live execution

Each question receives full execution and actual cancellation in both panels. The ten models, panel priorities, prompt format, parser, greedy decoding, and 128-token limit are unchanged. Capacity probes use validation prompts, with test prompt lengths informing the memory bound. Panel A runs batches of 151 questions followed by a final batch of 115; Panel B runs batches of 128 followed by 20. Neither final batch is padded with repeated questions.

The execution yields 18 matched workload pairs in one session. Panel A contributes eight pairs and Panel B ten, with balanced randomized arm order within each panel. Both arms share question identities and batch sizes. The package contains 192 validation traces, 2,344 test full traces, and 2,344 actual-cancellation records, for 4,880 task-arm records and 24,400 model-task records. Calibration is frozen before either panel begins testing. All 11,720 model-task sequences under cancellation match a token prefix of their paired full sequences.

K.3 Reference fidelity and actual resource savings

Both panels save generated work, with the combined correct count unchanged at 1,899. Each panel preserves 1,171 of 1,172 reference decisions. Panel A changes one correct answer to an incorrect answer; Panel B corrects one incorrect answer. All 2,344 live decisions match the separately reconstructed closure action on their paired reference traces. Table 42 reports correct-answer counts as well as resource use, keeping reference fidelity distinct from task accuracy.

Table 42: Complete-test accuracy and live resources. Token counts are means per question; correct-answer counts cover the full test. Time saving uses paired workload totals.

Panel	Full correct	Stop correct	Full tokens	Stop tokens	Token saving	Time saving
A	885	884	156.99	50.53	67.8%	10.8%
B	1014	1015	179.97	88.37	50.9%	4.3%

The savings are also present on the 916 previously unused questions. Generated tokens fall by 68.2% in Panel A and 50.8% in Panel B, with 915 questions retaining both reference decisions. The previously used group also reduces tokens in both panels and preserves both references on 255 of 256 questions. Table 43 gives exact correct-answer and reference-match counts for each group.

Every matched workload pair favors cancellation in wall time. Panel A requires 186.12 seconds under cancellation versus 208.72 seconds under full execution; Panel B requires 193.23 versus 201.99 seconds. Table 44 includes all 18 pairs and their arm order. These measurements describe the executed workloads within one session. They are not per-query latency estimates or independent deployment replications. Because batches mix history groups, their wall time cannot be assigned separately to the previously unused questions.

Table 43: Test history groups. Token saving is relative to each group's full execution; mixed-batch wall time is not attributed to subgroups.

Group	Panel	Questions	Reference matches	Full / stop correct	Token saving
Unused	A	916	916/916	689/689	68.2%
Used	A	256	255/256	196/195	66.5%
Unused	B	916	915/916	793/794	50.8%
Used	B	256	256/256	221/221	51.1%

Table 44: Every paired test workload. Order indicates the first arm. Wall times cover whole workloads within one runtime session.

Panel	Tasks	First arm	Full (s)	Stop (s)	Token saving	Time saving
B	128	Stop	17.89	17.40	48.7%	2.7%
B	128	Stop	23.21	21.87	51.2%	5.7%
B	128	Stop	23.14	22.47	52.3%	2.9%
B	128	Full	21.17	19.90	54.4%	6.0%
B	128	Full	20.47	20.07	49.6%	2.0%
B	128	Stop	22.69	22.04	52.0%	2.9%
B	128	Full	22.33	22.25	47.6%	0.4%
B	128	Full	20.76	17.69	50.1%	14.8%
B	128	Stop	24.45	24.22	52.6%	0.9%
B	20	Full	5.88	5.32	49.0%	9.5%
A	151	Full	26.72	22.72	69.4%	15.0%
A	151	Stop	26.42	22.98	69.1%	13.0%
A	151	Full	29.88	25.63	69.6%	14.2%
A	151	Stop	25.51	24.95	63.6%	2.2%
A	151	Stop	25.08	22.95	68.6%	8.5%
A	151	Full	27.95	23.63	68.4%	15.5%
A	151	Full	27.26	25.24	67.4%	7.4%
A	115	Stop	19.90	18.04	65.9%	9.4%

K.4 Prefix contribution and the deferral rule

Prefix observations add substantial savings beyond exact completed votes. On the same 2,344 uninterrupted reference traces, completed-vote closure saves 34.6% of tokens for plurality, while prefix-informed global closure saves 58.6%. The former preserves every reference decision, and the latter preserves 2,343 of 2,344. The 24.0-percentage-point difference measures the additional saving from prefix domains without selective freezing or further model calls.

The same domains support a fixed decision rule that can defer. Strict majority returns a label only when at least three of the five terminal outputs agree, and otherwise returns REVIEW. Completed-vote closure saves 32.4% of tokens, whereas prefix closure saves 57.9%, preserving all 2,344 reference actions. Both retain the reference deferral rate of 14.5%. Table 45 gives the complete fixed-trace comparison.

Table 45: Complete-test fixed-trace comparison. Each stopping policy is compared with its own full-computation decision.

Rule	Policy	Token saving	Reference matches	Deferral
Majority	Completed votes	32.4%	2,344/2,344	14.5%
Majority	Prefix domains	57.9%	2,344/2,344	14.5%
Plurality	Completed votes	34.6%	2,344/2,344	0.0%
Plurality	Margin corrected	2.0%	2,344/2,344	0.0%
Plurality	Margin plug-in	61.5%	2,308/2,344	0.0%
Plurality	Prefix domains	58.6%	2,343/2,344	0.0%

The native-margin diagnostics retain the tradeoff seen on OpenBookQA. The plug-in adaptation saves 61.5% of tokens with 36 changed reference decisions, while its corrected form saves 2.0% with no changes. These are the same local adaptations described in Appendix I. Their results characterize those adaptations; they do not rank the official end-to-end MARS or Parallel-Probe systems.

K.5 Coverage and the changed decisions

Raw domains exclude a terminal mark on 16 questions, and two change the complete controller's decision. Table 46 lists every affected panel execution. On test row 425, Panel B's full terminal marks are Qwen3: D, Gemma: C, Qwen2.5: B, Phi-4: B, and OLMo-2: abstention. Full plurality selects B. At the first review, Gemma instead has prefix D and raw domain $\{D\}$, which excludes its terminal C.

Table 46: All raw coverage-miss identities. Repeated pending states are counted in the last column; task failures merge them.

Test row	Panel	Model	Prefix	Terminal	Group	States
30	B	Gemma-2-2B	D	C	Unused	1
287	B	Gemma-2-2B	D	A	Unused	2
316	B	Gemma-2-2B	C	A	Unused	3
425	B	Gemma-2-2B	D	C	Unused	2
431	B	Gemma-2-2B	D	A	Unused	2
495	A	Mistral-7B-v0.3	A	B	Unused	3
545	B	Gemma-2-2B	C	B	Unused	2
636	B	OLMo-2-7B	C	D	Unused	2
759	B	Gemma-2-2B	B	A	Unused	1
834	B	Qwen2.5-3B	A	D	Unused	1
856	B	Gemma-2-2B	C	A	Used	1
884	A	Granite-3.3-2B	D	C	Used	2
972	B	Qwen2.5-3B	B	D	Used	3
1059	B	OLMo-2-7B	D	C	Used	2
1061	B	Qwen2.5-3B	A	D	Unused	3
1108	B	Gemma-2-2B	B	A	Unused	1

Freezing preserves that incorrect singleton after Gemma is cancelled. The first completion is OLMo-2's abstention; the controller freezes Gemma's $\{D\}$ and Phi-4's $\{B\}$, then retains Qwen3 and

Qwen2.5. Their terminal D and B produce a tie under the frozen domains, resolved to D by Qwen3's priority. D is the correct answer, so the change raises Panel B's correct count from 1,014 to 1,015. It remains a reference-fidelity error: the objective is to preserve the full decision, whether that decision is correct or incorrect. Global prefix closure without freezing recovers B on this trace.

The second disagreement occurs on row 884 in Panel A and is already present under global prefix closure. As in the extension case, Granite's domain is $\{D\}$ although its terminal answer is C. The restricted domains select B, while the full tuple gives a C majority. C is correct, reducing Panel A's correct count from 885 to 884. The two changed decisions have opposite accuracy effects and the same failure of domain coverage.

The case illustrates the same coverage condition as the OpenBookQA disagreement. Coordinate invariance is calculated exactly within the current domains, but Gemma's true terminal value has already left their product. The premise of Theorem 1 is therefore absent. No risk level, quantile, parser, or test selection was changed in response to these outcomes. Across both complete splits, the resource benefit recurs while the recorded domain misses identify where reference fidelity can fail.

L Weighted Voting and Probability-Based Stopping

Decision closure also saves work when models have unequal voting weights. We test this case on the complete OpenBookQA and ARC-Challenge traces, changing the fixed aggregation rule while retaining the prefix domains. Selective freezing saves 52.6% to 68.4% of generated tokens across the four source-panel combinations. This experiment separates two choices: how models contribute to the full decision, and when unfinished outputs can no longer change that decision.

L.1 Weights fitted on validation answers

Model weights are estimated from the same 32 validation-fit questions used for each source. Let $c_{s,i}$ be the number answered correctly by model i on source s. Under a Bernoulli correctness model and a uniform $\mathrm{Beta}(1,1)$ prior, the posterior is $\mathrm{Beta}(c_{s,i}+1, 33-c_{s,i})$. Its mean gives

$$w_{s,i} = \frac{c_{s,i}+1}{34}, \qquad D_w(y) = \arg\max_{a\in\mathcal{A}} \sum_i w_{s,i}\mathbf{1}\{y_i = a\}. \tag{18}$$

Here, $\mathcal{A}$ contains the valid answer labels. Abstention contributes no vote. Ties use the existing model priority, and an all-abstaining panel returns the first valid label. All weights share the same denominator, so integer coefficients $c_{s,i}+1$ yield exactly the same decision. Table 47 gives every fitted coefficient.

The weights remain fixed throughout each test trajectory. Calibration and test answers do not select them, and neither the domain estimator nor its risk level changes. Given the fit data, D_w is a fixed deterministic map, so Theorem 1 applies under the same coverage condition. These accuracy-based weights do not model dependence among errors, so we evaluate their effect on full-panel accuracy separately from stopping fidelity.

L.2 Complete-trace protocol and reference accuracy

The replay retains every question and completion snapshot from both complete tests. There are 1,672 questions and 3,344 panel trajectories, including the three- and five-option ARC questions. Counts, source quantiles, parsing, tie priorities, and history groups are those of Appendices J and K. The weighting formula and replay policies are specified before test generation; their evaluation uses the resulting full traces without new model calls.

Each stopping policy is compared with the complete decision under its own voting rule. Completed-vote closure uses unrestricted domains for pending calls. Prefix-global closure uses the nested domains and stops the entire panel when its decision image is a singleton. Selective freezing also stops irrelevant calls before global closure. A frozen call retains its last domain and token count; its later completion in the uninterrupted trace does not trigger a review. Generated work is summed at these stopping points. Token accounting excludes new controller overhead. Physical timing remains measured for the original live rules.

Table 47: Validation-fitted coefficients. Each weight is the integer coefficient divided by 34; all weights are fixed during testing.

Source	Panel	Model	Fit tasks	Correct	Coefficient
OpenBookQA	A	SmolLM2-1.7B	32	17	18
OpenBookQA	A	OLMo-2-1B	32	8	9
OpenBookQA	A	Granite-3.3-2B	32	22	23
OpenBookQA	A	Phi-3.5-mini	32	27	28
OpenBookQA	A	Mistral-7B-v0.3	32	27	28
OpenBookQA	B	Qwen3-4B	32	23	24
OpenBookQA	B	Gemma-2-2B	32	21	22
OpenBookQA	B	Qwen2.5-3B	32	26	27
OpenBookQA	B	Phi-4-mini	32	25	26
OpenBookQA	B	OLMo-2-7B	32	23	24
ARC-Challenge	A	SmolLM2-1.7B	32	16	17
ARC-Challenge	A	OLMo-2-1B	32	6	7
ARC-Challenge	A	Granite-3.3-2B	32	22	23
ARC-Challenge	A	Phi-3.5-mini	32	30	31
ARC-Challenge	A	Mistral-7B-v0.3	32	24	25
ARC-Challenge	B	Qwen3-4B	32	20	21
ARC-Challenge	B	Gemma-2-2B	32	21	22
ARC-Challenge	B	Qwen2.5-3B	32	27	28
ARC-Challenge	B	Phi-4-mini	32	26	27
ARC-Challenge	B	OLMo-2-7B	32	24	25

Unequal weights improve the full decision most in Panel A. Table 48 compares both full rules with weighted selective freezing. The weighted reference gains 25 correct OpenBookQA answers and 73 correct ARC answers in Panel A. Panel B retains its OpenBookQA count and loses two correct ARC answers. Closure retains the weighted-reference correct counts in the first three combinations and adds one correct answer in the fourth. That correction still counts as a reference disagreement.

Table 48: Accuracy under each complete rule and weighted selective closure. Changed correct counts do not redefine reference fidelity.

Source	Panel	Equal full	Weighted full	Weighted stop	Reference matches
OpenBookQA	A	368	393	393	500/500
OpenBookQA	B	424	424	424	499/500
ARC-Challenge	A	885	958	958	1,172/1,172
ARC-Challenge	B	1014	1012	1013	1,171/1,172

L.3 Resource savings and prefix contribution

Prefix domains reduce work beyond what completed votes alone can determine. Under weighted voting, completed-vote closure saves 14.8% to 44.2% of tokens across the four combinations. Prefix-global closure raises the reduction in every combination, to 50.4% through 69.8%. Selective freezing provides a further reduction in three combinations (Table 49). On ARC Panel A, its 68.0% saving is 1.8 percentage points below global closure. Freezing removes a future review event along with the call, so the remaining calls can run longer. Prefix constraints supply the main saving, while the additional benefit of freezing depends on the review schedule.

The savings also hold on questions unused in earlier development or evaluation. All 96 previously used OpenBookQA questions and all 256 previously used ARC questions retain both weighted references. Among previously unused questions, the corresponding counts are 403 of 404 and 915 of 916. Table 50 reports both groups without treating two executions of one question as independent examples.

Both weighted-reference disagreements occur outside the coverage condition. Table 51 identifies them. Weighted prefix-global closure and selective freezing change the same references to the same labels, so these disagreements are already present without selective cancellation. Every visited-

Table 49: Weighted-vote token savings on fixed full traces. Freezing changes the review schedule as well as generated work.

Source	Panel	Completed votes	Prefix-global	Selective
OpenBookQA	A	22.3%	65.8%	68.4%
OpenBookQA	B	44.0%	50.4%	52.6%
ARC-Challenge	A	14.8%	69.8%	68.0%
ARC-Challenge	B	44.2%	51.9%	54.7%

Table 50: Weighted reference fidelity by history group, merging the two panel executions per question.

Source	Group	Questions	Both references	A token saving	B token saving
OpenBookQA	Used	96	96/96	68.3%	51.3%
OpenBookQA	Unused	404	403/404	68.4%	52.9%
ARC-Challenge	Used	256	256/256	66.9%	55.0%
ARC-Challenge	Unused	916	915/916	68.3%	54.7%

domain case that contains the full terminal tuple preserves its weighted reference. This behavior is consistent with the theorem's separation of exact decision computation from statistical domain coverage.

Table 51: Every weighted-reference change. Prefix-global and selective closure give the same changed labels, outside visited-domain coverage.

Source	Test row	Panel	Reference	Stopped	Truth
OpenBookQA	293	B	D	C	A
ARC-Challenge	1061	B	D	A	A

L.4 A PROBABILITY-THRESHOLD COMPARISON

A continuous probability score offers a different stopping criterion. We evaluate one explicit construction using the existing categorical estimates. At history h, a completed call has a point mass at its observed output; a pending call has the probability vector from Equation 4. Denote these marginals by $\widetilde{p}_i(\cdot \mid h)$. For either fixed voting rule, a product approximation gives the decision probability

$$\widetilde{P}_h(a) = \sum_{y:D(y)=a} \prod_i \widetilde{p}_i(y_i \mid h). \tag{19}$$

The policy stops globally when $\max_a \widetilde{P}_h(a) \geq 0.95$ and returns the most probable label. Equal probabilities use the canonical label order. We fix the threshold before this replay at 0.95, using the original 0.05 risk target. At full completion, the distribution reduces to the exact reference decision.

The product approximation adds a statistical condition that exact decision images do not require. The fitted marginals alone do not establish their conditional independence or the accuracy of their joint decision probability at the stopping time. Thus this score does not inherit the trajectory-level conformal guarantee (Vovk et al., 2005; Angelopoulos & Bates, 2023). Table 52 compares it with prefix-global closure, so both policies use global stopping on the same full traces.

Exact prefix-global closure saves more tokens than probability stopping in three of the four weighted combinations, with the same disagreement counts. On weighted ARC Panel B, probability stopping saves 56.2% with six reference changes, compared with 51.9% and one change for prefix-global closure. Every observed disagreement fraction is below 5%. The comparison identifies a workload-dependent tradeoff within this fixed probability model.

Probability-based stopping and cost-sensitive control answer different questions. The score in Equation 19 estimates agreement with a fixed full decision. A controller that trades answer quality against delay and computation must also estimate future decision value and remaining resource use. Section 7 explains why review timing and physical execution enter that broader objective.

Table 52: Global stopping under exact prefix domains and the fixed probability threshold. Each rule uses its own reference.

Source	Panel	Rule	Exact saving	Changes	Prob. saving	Changes
OpenBookQA	A	Equal	68.7%	0	62.5%	2
OpenBookQA	A	Weighted	65.8%	0	56.8%	0
OpenBookQA	B	Equal	48.7%	0	48.1%	0
OpenBookQA	B	Weighted	50.4%	1	49.9%	1
ARC-Challenge	A	Equal	68.9%	1	63.8%	8
ARC-Challenge	A	Weighted	69.8%	0	60.5%	0
ARC-Challenge	B	Equal	49.6%	0	49.6%	0
ARC-Challenge	B	Weighted	51.9%	1	56.2%	6

M Execution Checks and Stress Tests

M.1 Raw-record and policy consistency

Two complementary checks connect the tables to the executed workloads. A separate raw-record pass reconstructs full plurality from terminal worker marks and verifies live decisions, task accuracy, token totals, and paired worker prefixes against the exported records. The execution program additionally compares its live controller with a separate fixed-trace policy implementation. Table 53 distinguishes these recorded checks from the empirical reference-fidelity endpoint.

Table 53: Recorded execution checks. Paired terminal decisions, accuracy, token totals and worker prefixes were additionally checked directly from raw records.

Study	Full traces	Live/shadow	Paired prefixes	Cancel observed/requested
Initial	384	384/384	1,920/1,920	923/924
Extension	1,024	1,024/1,024	5,120/5,120	2,407/2,407
OpenBookQA	1,000	1,000/1,000	5,000/5,000	3,102/3,102
ARC-Challenge	2,344	2,344/2,344	11,720/11,720	7,566/7,566

The raw-record checks agree across all four studies. Every generated sequence in a cancellation arm is a prefix of its paired full sequence. This supports interpreting the physical intervention as removing suffix computation while preserving the observed prefix. The fixed-trace implementation also reports no reference disagreement on a visited path whose domains contain its terminal tuple. Bootstrap intervals and exhaustive enumeration are produced by the reproduction program; the additional raw-record pass does not repeat those computations.

M.2 Exhaustive finite-state checks

The proof gives generality, while exhaustive enumeration tests the implementation of its finite set operations. Table 54 reports two complete map families. The first includes every binary-output map on three binary coordinates. The second includes every ternary-output map on two ternary coordinates.

Table 54: Exhaustive finite-state sanity checks for arbitrary deterministic maps. No violation was found.

Map family	Maps	Domain states	Persistence checks	Fidelity checks
Three binary coordinates	256	6,912	51,216	30,224
Two ternary coordinates	19,683	964,467	3,169,422	4,992,219
Total	19,939	971,379	3,220,638	5,022,443

The enumeration also performs 895,644 simultaneous-freezing checks. It automatically recovers the expansion failure represented in Table 5. These tests do not extend the empirical claim to new

decision maps; they check that the general proof object was implemented without a plurality-specific shortcut.

M.3 Runtime intervention audit

The live result requires evidence that cancellation happened and that retained streams preserved their semantics. All 3,840 initial-study worker records ended without error. Every prefix stream matched its worker output token for token. Workers observed 923 of 924 cancellation requests, with at most one token step after observation. Same-trace and live-to-shadow decisions both matched in all 384 trajectories.

Decoder checks precede scientific generation. Each of the ten models produced token-identical greedy outputs across two repeated runs on a held-out diagnostic prompt. This check uses the current tokenizer and decoder, without requiring agreement with a stored prompt length or answer. The diagnostic question is excluded from the fit, calibration, and confirmation roles.

The capacity probe selects concurrency by measured throughput within a memory bound derived from model weights, prompt allocations, and key-value cache growth. For the initial test it selected 91 queries in Panel A and 84 in Panel B. The balanced scientific clusters still contain 48 tasks, so they do not fill either selected capacity. The extension and complete tests use their own measured capacities, reported in the corresponding protocols. These workloads test batched cancellation; they do not establish maximal hardware utilization.

M.4 Arm-order and sampling stress tests

Low-load arm order is balanced within every benchmark source. Both order halves favor closure for tokens and wall time in both panels. Batched clusters use opposite orders, and all eight cluster effects favor closure. These checks reduce concern about warm-cache or first-arm effects, while the small number of batched clusters remains the dominant uncertainty there.

Leave-one-source-out task summaries preserve zero decision disagreements in all six removals. Removing OpenBookQA also removes the only coverage failure; removing any other source leaves one failure among 80 tasks. These summaries are sensitivity checks, not alternative registered endpoints, because the source-specific conformal construction and the pooled task unit were fixed before testing.

M.5 Scope of the results

The theorem covers finite deterministic maps; the experiments test three voting rules on language model outputs. Table 55 summarizes the assumptions and the evidence needed to extend these results.

Table 55: Supported scope and evidence required for broader claims.

Object	Supported here	Needed for extension
Decision map	Any fixed finite deterministic map, pathwise; plurality, weighted voting, and majority with deferral, empirically	New analysis for stochastic or changing maps
Terminal domains	Source-wise split-conformal domains over registered prefix states	Representative calibration for a new source or deployment distribution
Language output	Parsed finite multiple-choice marks plus abstention	A finite structured representation for open-ended generation
Execution semantics	Ten models in two panels, with complete live-to-shadow agreement	Intervention tests when cancellation changes other calls, shared state, or tool interactions
Resource effect	Two low-load studies and repeated batched workloads	Independent sessions for deployment-level batched estimates
Control objective	Reference-decision fidelity and generated-work reduction	A review-time or Bellman model for latency-optimal waiting

Within the evaluated setting, the decision is often fixed before all streams finish, and the controller identifies that event without observing future marks. The additional evidence in the last column would support broader deployment claims.

N Reproduction Details

N.1 Execution order

One standalone Colab cell executes the entire evidence pipeline from public inputs. It downloads the models and datasets, generates the development trajectories, fits the shared categorical estimator, and calibrates its source quantiles. It then executes the initial confirmation and independent extension. Separate validation-only fitting and calibration precede the two complete-test evaluations. Finally, the same program computes alternative-rule replays, statistical summaries, finite-state checks, tables, and numerical figure data from its newly generated outputs.

Runtime configuration appears at the top of the cell. Scientific defaults fix model and dataset revisions, row selections, panel priorities, prompts, decoding, risk level, and randomization seeds. Fitted counts, quantiles, voting weights, predictions, and performance are computed during the run. No earlier result archive or frozen fitted estimator is required. Temporary model and dataset caches remain on runtime storage; completed paired slices and the final result archive can persist in Drive.

The extension freezes the estimator fitted earlier in the same run. Its schedule covers 48 new questions per source, a balanced low-load subset, and at least eight high-load workload pairs per panel. Repeated batch positions are retained as executions, not new questions. Each worker is indexed by cohort, panel, load, workload, task, and model so that batched records remain distinct.

Weighted-rule analysis uses the new validation-fit answers and uninterrupted test traces produced in the same run. It fits weights independently of calibration and test outcomes, then evaluates each stopping rule against its own complete decision. No extra model calls are needed for this replay. Per-task records keep reference actions, stopped actions, accuracy, token counts, coverage, and history groups separate.

N.2 Artifact schema

The final archive separates design and provenance, worker, arm, domain, summary, and method-check records. Together they preserve the protocol, prefix streams, cancellation events, controller states, resource endpoints, and failure cases. Exact repository revisions and package versions remain in the provenance record. Completed paired slices are reused on resume; an unfinished pair is repeated together. A workload spanning multiple sessions is explicitly identified rather than reported as one uninterrupted deployment.

N.3 Claim-to-evidence correspondence

Every central claim has a distinct evidence path. Table 56 is also a guide for checking a future reproduction.

N.4 Reproduction checks

Trace replay and new physical execution have different targets. Replaying the archive should recover its complete-panel decisions, raw and nested domains, and policy outputs exactly. A new live run can change completion order, observed prefixes, and token totals even under greedy decoding, because reviews occur at completion events. It should report its own coverage, fidelity, and paired resource effects. Decoder and public artifact revisions also form part of the generation condition.

The extension section of the archive contains 288 new questions and 10,240 worker records. It yields 1,024 reference trajectories including repetitions, one question with a decision disagreement, and four questions with raw coverage misses. Its primary rule replay retains 576 trajectories, one per question and panel at high load. These counts keep repeated execution distinct from independent question coverage.

Table 56: Claim-to-evidence map for the manuscript.

Claim	Formal support	Empirical or executable support
Singleton image preserves the full decision	Theorem 1	5,022,443 finite-state fidelity checks and 384 fresh same-trace matches
Irrelevant coordinates may be frozen together	Proposition 1	895,644 finite-state freezing checks and all live relevance histories
Nested domains preserve earlier irrelevance	Proposition 2	3,220,638 persistence checks and no relevance re-entry after freezing
Prefix domains have a finite-sample risk interface	Propositions 3 to 5	Initial raw coverage: one failure among 96 tasks, with a 4.846% upper bound
Closure preserves the evaluated endpoint	Coverage implication plus fixed plurality	Initial: 96/96 tasks; extension: 287/288; OpenBookQA: 499/500; ARC-Challenge: 1,170/1,172; all live-to-shadow decisions agree
Closure reduces live resources	No theorem claim	Two paired low-load studies, repeated high-load workloads, and complete-test batched pairs
Prefix domains drive most gain beyond completed votes	Decision-image decomposition	Four-policy replay and paired task-bootstrap contrasts
The domains support deferral	Fixed strict-majority map	Prefix-global saving: extension 43.1%, OpenBookQA 56.5%, ARC-Challenge 57.9%; reference matches 576/576, 998/1,000, and 2,344/2,344
Closure supports unequal voting weights	Theorem 1 with fixed D_w	Two full-test replays save 52.6% to 68.4% of tokens; 1,670 of 1,672 task groups preserve both weighted references

The complete OpenBookQA study contains 96 validation questions and all 500 test questions: 192 validation traces, 1,000 test reference traces, and 1,000 cancellation records, totaling 10,960 model-task records. Its validation quantile is 0.907407 at rank 62. The test has 999 reference matches across 1,000 panel executions, five questions with raw coverage misses, and six workload pairs. Strict-majority prefix-global closure preserves 998 of 1,000 actions. These outcomes use their own validation-fitted estimator.

The ARC-Challenge study contains 192 validation traces, 2,344 test reference traces, and 2,344 cancellation records, totaling 24,400 model-task records. Its validation quantile is 0.827243 at rank 62. There are two question-level reference disagreements, sixteen questions with raw coverage misses, and eighteen matched workload pairs. Prefix-global closure preserves 2,343 plurality references and all 2,344 majority-with-deferral actions. Appendix K reports the complete numerical results.

The two full-test studies supply 3,344 reference trajectories for both equal and weighted voting. Four stopping policies give 26,752 rule-policy records, excluding the complete-reference rows and other diagnostics. Equal-vote selective decisions match all 3,344 live decisions. Under weighted voting, prefix-global and selective closure change one OpenBookQA and one ARC-Challenge reference, with no recorded disagreement on a covered visited path. These fixed-trace outcomes are separate from the physical execution checks above.

Models and datasets retain their original licenses and access terms. The artifacts redistribute no weights or benchmark corpora; shared code, public row indices, prompts, derived outputs, controller states, and aggregate results remain subject to those terms.