COMPUTE-OPTIMAL STOPPING FOR MULTI-LLM DELIBERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-LLM systems can spend substantially different amounts of inference compute on the same query, making adaptive stopping essential for balancing answer quality and computational cost. We present DWELL, a cost-sensitive stopping framework for selecting among fixed-depth multi-LLM deliberation policies. Under diminishing marginal accuracy per unit cost on the efficient frontier, we show that the cheapest utility-maximizing endpoint follows a threshold rule and that dominated depths can be excluded as stopping endpoints. In a controlled five-seed simulation, the selected depth-four endpoint achieves 0.950 ± 0.011 accuracy at a normalized cost of 7.1, compared with 0.964 ± 0.010 accuracy at a cost of 19.2 for mixture-of-agents. The resulting operating point is $2.7\times$ cheaper while sacrificing only 1.4 accuracy points. A lightweight confidence model obtains an expected calibration error of 0.102, and our conditional headroom bound connects calibrated confidence to the value of further deliberation. A held-out 60-item ARC-Challenge study with three API-served workers further characterizes how the preferred endpoint changes with the cost coefficient λ. Together, these results establish a principled mechanism for selecting deliberation depth under explicit accuracy–cost trade-offs while separating controlled evidence from the preliminary live-model evaluation.

1 INTRODUCTION

Multi-LLM deliberation trades inference cost for answer quality by purchasing workers or deeper aggregation. Coordination research selects models, roles, instructions, and topology (Xu et al., 2026; Nielsen et al., 2026; Tang et al., 2026; Chen et al., 2025); routing uses query and capability evidence (Zhang et al., 2026b; Ratnakar et al., 2026); self-consistency, debate, consensus, and layered aggregation prescribe interaction patterns (Wang et al., 2023b; Du et al., 2023; Chen et al., 2024; Wang et al., 2024). We ask which endpoint maximizes accuracy minus cost for a stated cost coefficient.

The highest-accuracy protocol need not maximize utility: additional workers incur cost, and rankings depend on the model pool and accounting convention (Li et al., 2024; Huang et al., 2025). An explicit analysis must remove dominated endpoints, measure marginal accuracy per unit cost, and resolve ties; otherwise a favorable global budget can be mistaken for adaptive orchestration.

DWELL selects among fixed-depth multi-LLM protocols using measured expected accuracy and cost. We evaluate coefficient-sensitive endpoint selection in a five-seed simulation and a preliminary ARC-Challenge study with three API-served models.

Distribution-level endpoint selection differs from query-adaptive control. Difficulty-aware scaling (Snell et al., 2024; Lin et al., 2025), learned quality–cost routing (Huang et al., 2026; Ong et al., 2025; Zhang et al., 2026b), and cascade deliberation (Chang et al., 2026) motivate our fixed-depth reference and conditional confidence criterion.

We therefore make four bounded contributions:

- We formulate cost-sensitive endpoint selection for a predefined distribution-level fixed-depth policy family and prove a cheapest-optimum threshold characterization under diminishing frontier returns.

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.

- We state a conditional-expectation headroom bound under state-conditional calibration, while identifying why aggregate ECE does not establish its premise or provide pointwise safety.
- We evaluate fixed-depth policies, oracle endpoint selections, confidence calibration, and representative multi-LLM deliberation protocols across explicit accuracy–cost operating points.
- We report ARC-Challenge results across four cost coefficients and analyze when cost-sensitive selection favors additional deliberation.

2 Related Work

Learned coordinators optimize model, role, instruction, or communication assignments (Xu et al., 2026; Nielsen et al., 2026; Zhang, 2026; Papamarkou et al., 2026). Graph routing and topology optimization address collaborator structure (Zhang et al., 2026b; Li et al., 2024; Chen et al., 2025; Luo & Shao, 2026), while reflection expands workflows (Shinn et al., 2023; Madaan et al., 2023; Tang et al., 2026). We instead analyze stopping over a fixed-depth policy family.

Test-time allocation depends on difficulty and can combine routing, reasoning, and precomputation (Snell et al., 2024; Lin et al., 2025; Huang et al., 2026). Planning, search, hints, and execution-grounded surprisal structure compute use (Wang et al., 2023a; Yao et al., 2023; Dong et al., 2026; Barnes, 2026). DWELL studies the distribution-level deliberation frontier.

Cost-aware routers estimate query–model quality and price (Ong et al., 2025; Ding et al., 2024; Xue et al., 2026; Zhang et al., 2026a; Kumaresan, 2026), including controllable reasoning, semantic gates, and cascades (Ding et al., 2025; Cao et al., 2026; Ratnakar et al., 2026; Chang et al., 2026). Benchmarks show that conclusions depend on the model pool and protocol (Huang et al., 2025), motivating explicit operating coefficients.

Marginal calibration does not establish the state-conditional premise needed for confidence-based stopping. LLM-as-a-judge studies likewise caution that evaluators are not error-free oracles (Zheng et al., 2023). We distinguish the exact frontier theorem from the conditional confidence proposition.

Self-consistency, search, debate, and mixture-of-agents define different endpoints (Wang et al., 2023b; Yao et al., 2023; Du et al., 2023; Wang et al., 2024). Diversity and topology alter those endpoints at a fixed depth (Chen et al., 2024; Feng et al., 2026; Li et al., 2024). Common cost accounting permits our distinct comparison: which observed endpoint maximizes accuracy minus cost?

3 The DWELL Framework

Figure 1 shows a sequential interface linking workers, answer synthesis, confidence, and stopping. A controller can observe the query, candidates, agreement, confidence, and budget. Only the logistic confidence block is instantiated in the experiments; the query-adaptive controller remains an unevaluated design target.

3.1 Analyzed Fixed-Depth Policy Family

The empirical object in this paper is a predefined family $\Pi_{\text{fix}} = \{\pi_1, \ldots, \pi_T\}$ of distribution-level fixed-depth policies. Policy π_t executes the first t stages of a prescribed worker order, applies the answer selector, and terminates at endpoint t. For a query $x \sim \mathcal{D}$, its returned answer is $\hat{y}_t(x)$, its expected accuracy is $A(t) = \mathbb{E}[\mathbb{1}[\hat{y}_t = y]]$, and its expected cumulative cost is $C(t)$. Selecting t after estimating these distribution-level quantities is an oracle model-selection operation rather than a per-query stopping policy.

The objective for a cost coefficient $\lambda > 0$ and budget B is

$$\mathcal{T}_B = \{t : C(t) \le B\},$$
$$t^\star(\lambda) = \arg\min_{t \in \mathcal{T}_B}{}^{\text{cheap}} \{-A(t) + \lambda C(t)\}. \tag{1}$$

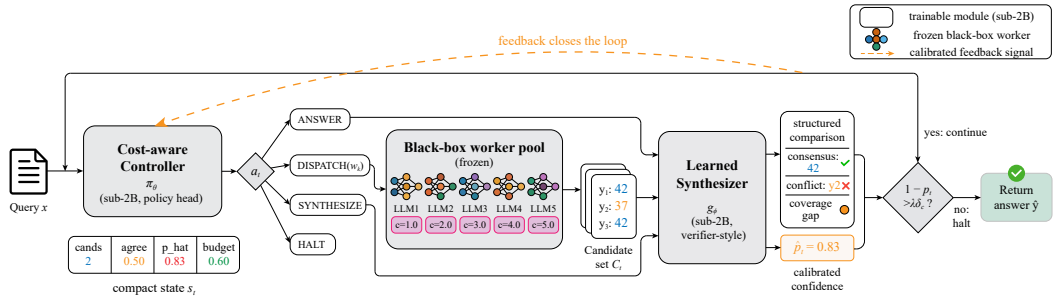


Figure 1: **DWELL separates candidate generation, answer synthesis, confidence estimation, and stopping.** The architecture defines a sequential interface in which a controller can purchase another black-box worker response or terminate with the current synthesized answer. The experiments in this paper evaluate prescribed fixed-depth prefixes of this interface and oracle endpoint selection; the query-adaptive controller loop (dashed arrow) specifies the design target for future per-query policies rather than an evaluated component. Worker costs and panel values shown are illustrative.

where $\arg\min^{\text{cheap}}$ breaks ties by lower cost, equivalently maximizing $U_\lambda(t) = A(t) - \lambda C(t)$. At least the cheapest endpoint must satisfy the budget. Sweeping λ selects among the same measured endpoints.

3.2 Answer and Halt Semantics

Each policy accumulates candidates $\mathcal{C}_t$ from the first t workers and selects among normalized answers by plurality. Both ANSWER and HALT terminate with that answer; distinguishing them permits an explicit synthesis request in the sequential interface. The state exposes agreement, confidence, used workers, and remaining budget.

Endpoint costs are not token-billing claims. Moving from t to t' costs $C(t') - C(t)$ and may require intermediate workers. A dominated depth is excluded as a stopping choice but may still be traversed, as with depth two below.

3.3 Implemented Confidence Model

A logistic head over candidate-panel features estimates endpoint correctness as $\hat{p}_t$. Its ECE is 0.102 over two thousand candidate sets. Proposition 1 relates state-conditional confidence to the maximum expected benefit of continuation.

Oracle selection in Eq. (1) uses aggregate $A(t)$ and $C(t)$ directly, independently of $\hat{p}_t$, separating calibration from frontier construction.

3.4 Query-Adaptive Controller Interface

The design interface exposes a controller state S_t containing query features, candidate summaries, $\hat{p}_t$, used workers, and remaining budget. An MLP or compact language-model policy would choose among dispatching an available worker, requesting synthesis, returning the current answer, and halting. The policy objective is the query-level analogue of Eq. (1),

$$J(\theta) = \mathbb{E}_{x,\tau\sim\pi_\theta}\left[\mathbb{1}[\hat{y}_\tau = y] - \lambda \sum_{a\in\tau} c(a)\right]. \tag{2}$$

This objective extends the fixed-depth criterion to trajectories whose stopping depth depends on the observed query state.

We do not train or evaluate this controller. The fixed-depth optimum and Proposition 1's conditional stopping criterion provide references for future query-adaptive policies.

4 COMPUTE-OPTIMAL ENDPOINT SELECTION

We analyze Π_{fix} under a common query distribution, with expected endpoint accuracy $A(t)$, cumulative cost $C(t)$, strictly increasing costs, and a budget-feasible cheapest endpoint. This is distribution-level model selection, not a query-conditioned MDP.

Definition 1 (Efficient stopping endpoints)**.** *An endpoint t is efficient if no cheaper endpoint u satisfies $A(u) \geq A(t)$. Write the budget-feasible efficient endpoints in increasing-cost order as $e_0, e_1, \ldots, e_m$, where e_0 is the cheapest feasible endpoint. For $j \geq 1$, define*

$$\rho_j = \frac{A(e_j) - A(e_{j-1})}{C(e_j) - C(e_{j-1})}. \tag{3}$$

An inefficient depth is dominated only as a stopping endpoint; executing a deeper policy may still pass through its computation. The frontier is thus a subset of admissible endpoints rather than a shortcut through the underlying execution sequence.

Assumption 1 (Diminishing frontier returns)**.** *The predefined fixed-depth frontier has positive, non-increasing marginal rates, $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_m > 0$. The assumption concerns distribution-level endpoint statistics and need not hold for an individual query. The measured frontier rates satisfy this ordering. Query-specific continuation decisions additionally depend on the state observed at each endpoint.*

Theorem 1 (Cheapest threshold optimum)**.** *Under Assumption 1, a maximizer of $U_\lambda(t) = A(t) - \lambda C(t)$ over the predefined budget-feasible fixed-depth family lies on the efficient endpoint frontier. Define the selected endpoint to be the cheapest maximizer of this utility. Let $\mathcal{J}_\lambda$ contain the feasible frontier indices reached only through strictly beneficial incoming slopes:*

$$\begin{aligned} \mathcal{J}_\lambda = \{j \in \{1, \ldots, m\} : C(e_j) \leq B, \\ \rho_\ell > \lambda \text{ for every } \ell \leq j\}, \\ j^\star_\lambda = \max(\mathcal{J}_\lambda \cup \{0\}), \\ t^\star_\lambda = e_{j^\star_\lambda}. \end{aligned} \tag{4}$$

Thus equality $\rho_j = \lambda$ does not justify purchasing step j, and a dominated depth is skipped only as the final stopping endpoint.

Proof. If endpoint t is dominated, there is a cheaper endpoint u with $A(u) \geq A(t)$. Therefore $U_\lambda(u) - U_\lambda(t) = A(u) - A(t) + \lambda(C(t) - C(u)) > 0$, so t cannot maximize utility. Along the efficient sequence,

$$U_\lambda(e_j) - U_\lambda(e_{j-1}) = \big(C(e_j) - C(e_{j-1})\big)(\rho_j - \lambda).$$

Under Assumption 1 the sign of these increments is non-increasing along the frontier, so utility increases while every incoming slope through endpoint e_j exceeds λ, is tied across an incoming step when $\rho_j = \lambda$, and decreases after the slopes fall below λ. The quantified condition in Eq. (4) selects the last endpoint reached entirely through strictly utility-increasing steps. Equality excludes the tied higher-cost endpoint and therefore implements the cheapest-maximizer convention. Feasibility of e_0 guarantees that the selection set is nonempty, while $C(e_j) \leq B$ enforces endpoint feasibility. □

Interpretation. Endpoint selection is global for a given λ. Excluding depth two as a stopping choice does not avoid its computation when reaching depth three.

Proposition 1 (Conditional headroom bound)**.** *Let Z_t denote the full stopping state at endpoint t, let $Y_t = \mathbb{1}[\hat{y}_t = y]$, and suppose a confidence function is state-conditionally calibrated, $\mathbb{E}[Y_t \mid Z_t] = p_t(Z_t)$. For any candidate continuation to endpoint u with incremental cost $\Delta C_{t,u} > 0$,*

$$\mathbb{E}[Y_u - Y_t \mid Z_t] \leq 1 - p_t(Z_t). \tag{5}$$

Consequently, if $1 - p_t(Z_t) \leq \lambda \Delta C_{t,u}$, the conditional expected utility improvement of that particular continuation is non-positive.

Proof. Because $Y_u \leq 1$, conditional expectation gives $\mathbb{E}[Y_u - Y_t \mid Z_t] \leq 1 - \mathbb{E}[Y_t \mid Z_t]$. State-conditional calibration replaces the final term by $p_t(Z_t)$ and yields Eq. (5). Subtracting $\lambda \Delta C_{t,u}$ from both sides proves the utility statement. The result is conditional in expectation and does not constrain the realized outcome of any individual query. □

Scope of the confidence result. Proposition 1 is a conditional-expectation statement rather than a pointwise guarantee. Its premise requires confidence calibrated with respect to the complete stopping state Z_t. The measured ECE of 0.102 summarizes marginal calibration and quantifies the remaining calibration gap. State-conditional calibration can be assessed more finely by stratifying reliability estimates over the controller's decision features.

5 Evaluation Setup

We evaluate a controlled five-seed fixed-depth simulation and a preliminary ARC-Challenge study (Clark et al., 2018) with Llama-3.1-8B, Qwen-2.5-7B, and GPT-4o-mini. The simulation varies worker skills, correlated errors, and normalized costs; ARC-Challenge uses live model outputs.

5.1 Fixed-Depth Sweep

The controlled simulation uses heterogeneous workers whose family-conditioned skill profiles create complementary candidate panels and whose errors include a shared correlation component. The five endpoints have cumulative normalized costs 1.0, 3.1, 4.7, 7.1, and 10.1, and each applies the same plurality answer selector over normalized candidates. Individual-worker performance is summarized by the best-single result of 0.827 ± 0.014 in Table 1. All five measured endpoints are feasible, so endpoint selection maximizes $A(t) - \lambda C(t)$ over the complete sweep for each λ; these results demonstrate controlled decision behavior and are not evidence of superiority over deployed LLM systems.

5.2 Comparison Protocols

The comparison includes best-single, self-consistency, multi-agent debate, mixture-of-agents, a cost-aware router, and a fixed-budget policy. Self-consistency resamples and votes, debate applies a fixed interaction pattern, and mixture-of-agents uses layered aggregation (Wang et al., 2023b; Du et al., 2023; Wang et al., 2024). The cost-aware router performs one-shot model selection following established routing formulations (Ong et al., 2025; Huang et al., 2025). The FixedBudget comparison selects the middle depth independently of λ, providing a direct test of whether one global depth remains optimal across cost regimes.

5.3 ARC-Challenge Evaluation

The ARC-Challenge study draws 120 examples, using 60 to configure the synthesizer and fit the confidence model, and a disjoint 60 for evaluation under random seed zero. Worker responses are requested once through OpenRouter at temperature 0.3 with a 256-token cap, reduced to a final multiple-choice letter, and cached so every policy sees the same outputs. Relative worker costs are 0.6, 0.8, and 1.0 for Llama-3.1-8B, Qwen-2.5-7B, and GPT-4o-mini, respectively. Reported protocol costs bill worker calls only: answer aggregation applies the same lightweight plurality selector to every protocol and is not billed. We evaluate $\lambda \in \{0.02, 0.06, 0.12, 0.20\}$ against best-single, three-sample self-consistency, one-layer mixture-of-agents, and two-round debate; because this is one held-out split rather than a repeated benchmark study, its comparisons are descriptive.

5.4 Metrics

We report exact-match accuracy and normalized mean inference cost, with repeated-run entries presented as mean $\pm$ standard deviation. Utility is defined as $U_\lambda = A - \lambda C$, allowing methods with different costs to be compared at a common operating coefficient. Confidence quality is measured with expected calibration error (ECE) over two thousand candidate sets and visualized with a reliability diagram. Frontier slopes are computed as the accuracy gain divided by incremental cost between adjacent efficient endpoints.

Table 1: **Selected depths expose the accuracy–cost trade-off.** Entries report mean ± standard deviation over five seeds; DWELL-ORACLE rows denote the fixed-depth endpoint selected for each cost coefficient.

Protocol	Accuracy	Cost
Best single	0.827 ± 0.014	2.4
Self-consistency	0.875 ± 0.014	12.0
Multi-agent debate	0.908 ± 0.013	8.4
Mixture-of-agents	0.964 ± 0.010	19.2
Cost-aware router	0.781 ± 0.025	2.5
Fixed budget	0.901 ± 0.009	4.7
DWELL-ORACLE (.01)	0.950 ± 0.011	7.1
DWELL-ORACLE (.025)	0.901 ± 0.009	4.7
DWELL-ORACLE ($\geq$ .03)	0.792 ± 0.023	1.0

5.5 ETHICS AND REPRODUCIBILITY

The study uses an existing benchmark and model-generated outputs and introduces no human-subject experiment. All comparisons apply the same accuracy metric and normalized cost accounting, and the reported endpoint values permit direct reconstruction of the utility comparisons; deployment-specific safety assessment remains outside the scope of this evaluation.

6 RESULTS

We compare accuracy–cost endpoints, marginal rates, calibration, and coefficient-dependent selections.

6.1 ACCURACY–COST VALUES

Table 1 compares accuracy and cost. Mixture-of-agents attains the highest accuracy but also the highest cost.

The DWELL-ORACLE rows select depths four, three, and one as λ increases. Depth four is 1.4 accuracy points below mixture-of-agents at $2.7\times$ lower cost.

Figure 2 plots protocol summaries and oracle selections on the same fixed-depth frontier; oracle selection adds no new operating points.

6.2 FRONTIER ASSUMPTION AND ENDPOINT TIES

Figure 3 shows efficient endpoints $\{1, 3, 4, 5\}$ with marginal rates 0.0295, 0.0204, and 0.0037, satisfying Assumption 1. Depth two is dominated as a stopping endpoint.

The threshold selects distribution-level depths four, three, and one at $\lambda = 0.01$, 0.025, and $\lambda \geq 0.03$. Ties retain the cheaper endpoint; excluding depth two as a stopping choice does not skip its computation.

6.3 DECISION BOUNDARIES AND SENSITIVITY

For any two observed endpoints i and j with $C(j) > C(i)$, their utility ordering changes at the empirical break-even coefficient

$$\lambda^{\times}_{i \to j} = \frac{A(j) - A(i)}{C(j) - C(i)}. \tag{6}$$

This coefficient is an interpretable exchange rate rather than a separately fitted hyperparameter. Table 2 reports the transitions implied by the observed means. Because the simulation frontier rates decrease, these local crossovers also determine the global endpoint sequence characterized by Theorem 1.

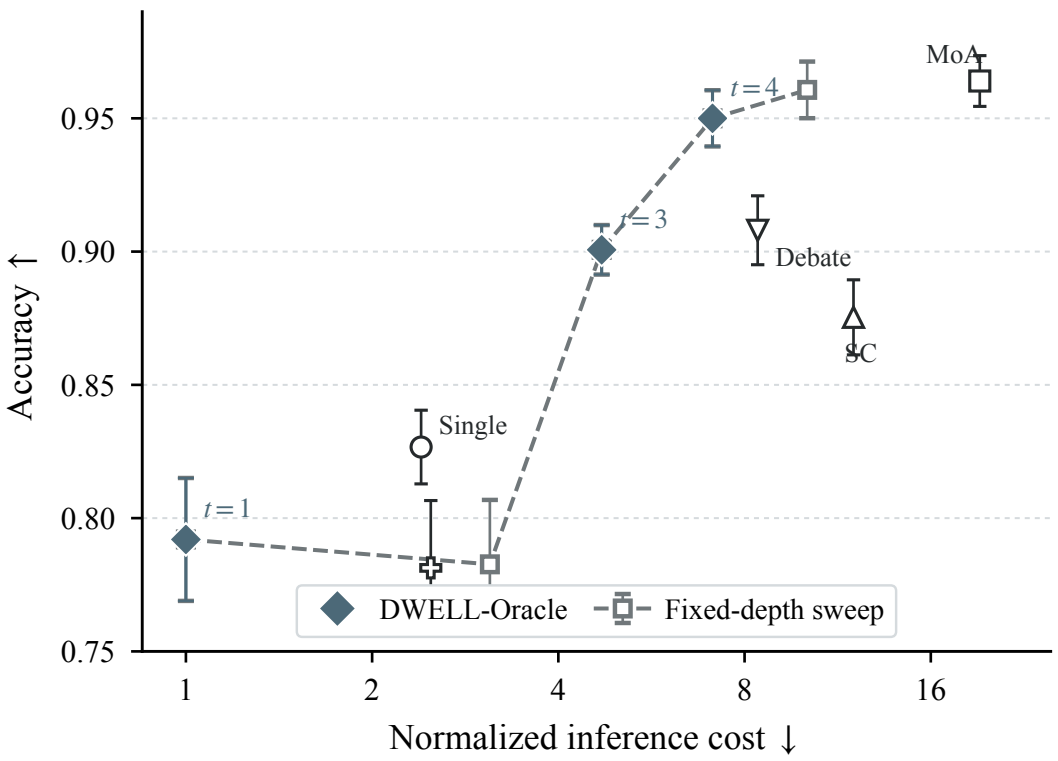


Figure 2: **Oracle selection returns compute-optimal sweep endpoints.** Accuracy and normalized inference cost are reported over five seeds; DWELL-ORACLE selects operating points from the fixed-depth frontier as λ changes.

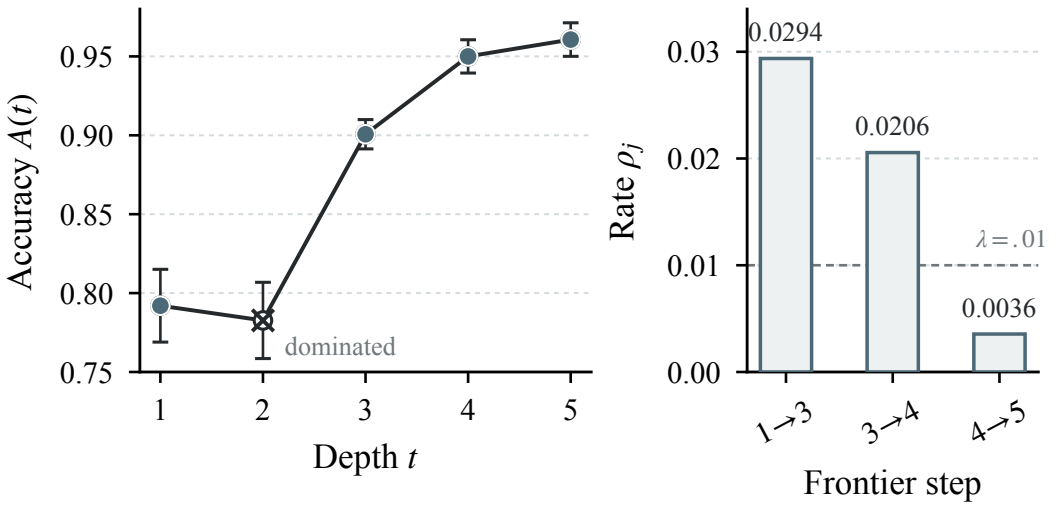


Figure 3: **Marginal frontier rates decrease with deliberation depth.** The left panel reports endpoint accuracy with standard-deviation error bars; the right panel shows rates $0.0295 \geq 0.0204 \geq 0.0037$. Depth two is excluded only as a stopping endpoint.

The controlled frontier therefore selects depth five for $\lambda < 0.0037$, depth four for $0.0037 \leq \lambda < 0.0204$, depth three for $0.0204 \leq \lambda < 0.0295$, and depth one for $\lambda \geq 0.0295$. On ARC-Challenge, self-consistency is dominated by the cheapest endpoint and debate is dominated by mixture-of-agents, leaving a break-even value of $\lambda^{\times} = 0.100/1.8 \approx 0.0556$ between the two nondominated

Table 2: **Observed utility crossovers identify stable selection regimes.** Values are computed from the reported mean accuracies and costs; ties select the cheaper endpoint.

Transition	ΔA	ΔC	$\lambda^{\times}$
Simulation: depth $1 \to 3$	.109	3.7	.0295
Simulation: depth $3 \to 4$	.049	2.4	.0204
Simulation: depth $4 \to 5$	.011	3.0	.0037
ARC: cheapest $\to$ MoA	.100	1.8	.0556

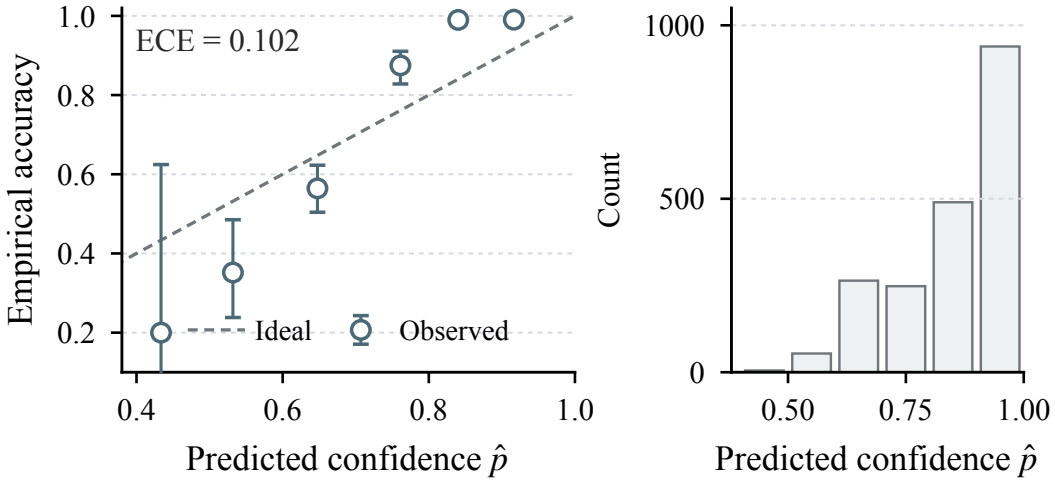


Figure 4: **The confidence model achieves ECE** 0.102**.** The left panel shows empirical accuracy with binomial 95% intervals over two thousand candidate sets; the right panel reports confidence-bin counts.

observations. At $\lambda = 0.02$, mixture-of-agents has a utility advantage of 0.064 over the cheapest endpoint; at $\lambda = 0.06$, the cheapest endpoint leads by 0.008. Because the evaluation uses a single 60-item held-out split, the observed crossover should be interpreted as a descriptive boundary rather than a statistically resolved ranking.

6.4 Confidence Calibration

Figure 4 reports ECE $= 0.102$ over two thousand candidate sets. The bin near 0.95 contains 939 cases, versus 5 near 0.45; the largest visible reliability gaps occur in sparse bins.

6.5 Fixed-Budget Comparison

A fixed budget selects depth three regardless of λ: it overspends when depth one is optimal and stops too early when depth four is preferred. One global depth therefore cannot optimize all cost coefficients.

Query-adaptive control could additionally condition on the query, using the frontier and confidence bound as references.

6.6 ARC-Challenge Results

In Table 3, best-single dominates self-consistency (equal 0.783 accuracy at costs 0.6 versus 1.8), and mixture-of-agents dominates debate (equal 0.883 accuracy at 2.4 versus 4.8). Dwell selects between the two efficient endpoints.

At $\lambda = 0.02$, mixture-of-agents has utility 0.835 versus 0.771 for best-single. The cheapest endpoint wins at $\lambda \in \{0.06, 0.12, 0.20\}$, consistent with the crossover $\lambda^{\times} \approx 0.0556$.

Table 3: **ARC-Challenge exposes coefficient-dependent endpoint preferences.** The DWELL rows select the utility-maximizing efficient endpoint at each λ; the break-even coefficient is $\lambda^{\times} \approx 0.0556$.

Protocol	λ	Acc.	Cost
Best single	–	0.783	0.6
Self-consistency	–	0.783	1.8
Mixture-of-agents	–	0.883	2.4
Multi-agent debate	–	0.883	4.8
DWELL	.02	0.883	2.4
DWELL	.06	0.783	0.6
DWELL	.12	0.783	0.6
DWELL	.20	0.783	0.6

7 DISCUSSION, LIMITATIONS, AND BROADER IMPACT

The theorem characterizes distribution-level endpoint selection, providing a reference for query-adaptive policies. The confidence bound additionally requires state-conditional calibration; measured aggregate ECE 0.102 does not establish that premise. Common cost accounting is essential when comparing different interaction structures (Wang et al., 2023b; Du et al., 2023; Wang et al., 2024; Ong et al., 2025).

Several limitations remain. The five-seed frontier is a controlled simulation and therefore validates the decision rule without establishing performance gains for real multi-LLM deployments. The ARC-Challenge study is limited to one 60-item held-out split, one random seed, and three worker models; its crossover at $\lambda^{\times} \approx 0.0556$ confirms that the preferred endpoint changes with λ in a live setting, but a larger repeated evaluation would be required to report confidence intervals on the boundary. The theorem additionally assumes diminishing marginal frontier rates, while marginal ECE may conceal calibration variation across task families, depths, or worker combinations.

Cost-aware orchestration can reduce unnecessary model calls, but premature stopping may lower reliability on difficult or safety-relevant queries. Applications should therefore set λ according to domain-specific error costs rather than inference price alone. Confidence estimates should also be monitored under distribution shift before they control consequential decisions.

8 REPRODUCIBILITY STATEMENT

The controlled-study figures use recorded per-condition means and standard deviations. Frontier rates, crossover values, and calibration intervals were computed deterministically from those records. ARC-Challenge fixes model identifiers, decoding settings, the seed split, and one response per worker–item pair before policy evaluation, so all policies use the same candidate panel and within-study API variation is controlled.

9 CONCLUSION

DWELL selects deliberation endpoints using an explicit accuracy–cost objective. Diminishing frontier returns yield a cheapest-optimum threshold; state-conditional calibration bounds continuation headroom. The selected simulation depth-four endpoint costs 37% of mixture-of-agents, and ARC-Challenge illustrates coefficient-dependent choices.

AI USE STATEMENT

AI tools were used for language polishing.

REFERENCES

Jarrod Barnes. Surprisal-guided selection: Compute-optimal test-time strategies for execution-grounded code generation, 2026. URL https://arxiv.org/abs/2602.07670.

Qi Cao, Shuhao Zhang, Ruizhe Zhou, Ruiyi Zhang, Peijia Qin, and Pengtao Xie. Models under scope: Scalable and controllable routing via pre-hoc reasoning, 2026. URL https://arxiv.org/abs/2601.22323.

Raeyoung Chang, Dongwook Kwon, Jisoo Lee, and Nikhil Verma. Cascadedebate: Multi-agent deliberation for cost-aware llm cascades. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 1329–1340. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-industry.93. URL http://dx.doi.org/10.18653/v1/2026.acl-industry.93.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.381. URL http://dx.doi.org/10.18653/v1/2024.acl-long.381.

Weize Chen, Jiarui Yuan, Chen Qian, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Optima: Optimizing effectiveness and efficiency for llm-based multi-agent system. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11534–11557. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.601. URL http://dx.doi.org/10.18653/v1/2025.findings-acl.601.

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, 2018. URL https://arxiv.org/abs/1803.05457.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid llm: Cost-efficient and quality-aware query routing, 2024. URL https://arxiv.org/abs/2404.14618.

Dujian Ding, Ankur Mallick, Shaokun Zhang, Chi Wang, Daniel Madrigal, Mirian Del Carmen Hipolito Garcia, Menglin Xia, Laks V. S. Lakshmanan, Qingyun Wu, and Victor Rühle. Best-route: Adaptive llm routing with test-time optimal compute, 2025. URL https://arxiv.org/abs/2506.22716.

Ziming Dong, Hardik Sharma, Evan O'Toole, Jaya Prakash Champati, and Kui Wu. Pay for hints, not answers: Llm shepherding for cost-efficient inference, 2026. URL https://arxiv.org/abs/2601.22132.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023. URL https://arxiv.org/abs/2305.14325.

Shangbin Feng, Wenxuan Ding, Alisa Liu, Zifeng Wang, Weijia Shi, Yike Wang, Shannon Zejiang Shen, Xiaochuang Han, Hunter Lang, Chen-Yu Lee, Tomas Pfister, Yejin Choi, and Yulia Tsvetkov. When one llm drools, multi-llm collaboration rules. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17048–17063. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.775. URL http://dx.doi.org/10.18653/v1/2026.acl-long.775.

Kaiyu Huang, Xingyu Wang, Mingze Kong, Zhubo Shi, Yuqian Hou, Hong Xu, Zhongxiang Dai, Minchen Yu, and Qingjiang Shi. Uniscale: Adaptive unified inference scaling via online joint optimization of model routing and test-time scaling, 2026. URL https://arxiv.org/abs/2605.30898.

Zhongzhan Huang, Guoming Ling, Yupei Lin, Yandong Chen, Shanshan Zhong, Hefeng Wu, and Liang Lin. Routereval: A comprehensive benchmark for routing llms to explore model-level scaling up in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3860–3887. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.208. URL http://dx.doi.org/10.18653/v1/2025.findings-emnlp.208.

Ramchand Kumaresan. Acar: Adaptive complexity routing for multi-model ensembles with auditable decision traces, 2026. URL https://arxiv.org/abs/2602.21231.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7281–7294. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.427. URL http://dx.doi.org/10.18653/v1/2024.findings-emnlp.427.

Kevin Lin, Charlie Snell, Yu Wang, Charles Packer, Sarah Wooders, Ion Stoica, and Joseph E. Gonzalez. Sleep-time compute: Beyond inference scaling at test-time, 2025. URL https://arxiv.org/abs/2504.13171.

Jingkai Luo and Yulin Shao. Cayley graph optimization for scalable multi-agent communication topologies, 2026. URL https://arxiv.org/abs/2604.09703.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651.

Stefan Nielsen, Edoardo Cetin, Peter Schwendeman, Qi Sun, Jinglue Xu, and Yujin Tang. Learning to orchestrate agents in natural language with the conductor, 2026. URL https://arxiv.org/abs/2512.04388.

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, 2025. URL https://arxiv.org/abs/2406.18665.

Theodore Papamarkou, Pierre Alquier, Matthias Bauer, Wray Buntine, Andrew Davison, Gintare Karolina Dziugaite, Maurizio Filippone, Andrew Y. K. Foong, Vincent Fortuin, Dimitris Fouskakis, Jes Frellsen, Eyke Hüllermeier, Theofanis Karaletsos, Mohammad Emtiyaz Khan, Nikita Kotelevskii, Salem Lahlou, Yingzhen Li, Fang Liu, Clare Lyle, Thomas Möllenhoff, Konstantina Palla, Maxim Panov, Yusuf Sale, Kajetan Schweighofer, Artem Shelmanov, Siddharth Swaroop, Martin Trapp, Willem Waegeman, Andrew Gordon Wilson, and Alexey Zaytsev. Position: agentic ai orchestration should be bayes-consistent, 2026. URL https://arxiv.org/abs/2605.00742.

Shivam Ratnakar, Abhiroop Talasila, and Vinayak K Doifode. Latentgate: Low-latency semantic routing via frozen-backbone probing of small language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 2284–2294. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-industry.153. URL http://dx.doi.org/10.18653/v1/2026.acl-industry.153.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366.

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

Yujin Tang, Edoardo Cetin, Jinglue Xu, Qi Sun, Stefan Nielsen, Vincent Richard, Haruto Goda, Iaroslav Tymchenko, Nhan Nguyen, Hyunin Lee, Mari Ashiga, Shashank Kotyan, So Kuroki, and Tarin Clanuwat. Sakana fugu technical report, 2026. URL https://arxiv.org/abs/2606.21228.

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

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models, 2023a. URL https://arxiv.org/abs/2305.04091.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023b. URL https://arxiv.org/abs/2203.11171.

Jinglue Xu, Qi Sun, Peter Schwendeman, Stefan Nielsen, Edoardo Cetin, and Yujin Tang. Trinity: An evolved llm coordinator, 2026. URL https://arxiv.org/abs/2512.04695.

Jiaqi Xue, Qian Lou, Jiarong Xing, and Heng Huang. R2-router: A new paradigm for llm routing with reasoning, 2026. URL https://arxiv.org/abs/2602.02823.

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, 2023. URL https://arxiv.org/abs/2305.10601.

Chenchen Zhang. Reinforcement learning for llm-based multi-agent systems through orchestration traces, 2026. URL https://arxiv.org/abs/2605.02801.

Wenbo Zhang, Lijinghua Zhang, Liner Xiang, and Hengrui Cai. Reasoning is not free: Robust adaptive cost-efficient routing for llm-as-a-judge, 2026a. URL https://arxiv.org/abs/2605.10805.

Zheyuan Zhang, Kaiwen Shi, Zhengqing Yuan, Zehong Wang, Tianyi Ma, Keerthiram Murugesan, Vincent Galassi, Chuxu Zhang, and Yanfang Ye. Agentrouter: A knowledge-graph-guided llm router for collaborative multi-agent question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 788–809. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.33. URL http://dx.doi.org/10.18653/v1/2026.acl-long.33.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.