Allocation, Not Volume: Test-Time Compute for Agentic Forecasting

Anonymous Authors[1]

Abstract

Test-time compute scaling has been studied extensively in verifiable domains such as math and code; how to spend an inference budget for forecasting future events, where no test-time verifier exists, is far less studied. We compare three multi-agent compute-allocation policies, static depth (*Predictor–Critic*), static breadth (*Ensemble*), and adaptive routing (*Hierarchical Orchestrator*), on a contamination-controlled benchmark of 228 label-balanced binary ForecastBench questions resolved strictly after every base model's knowledge cutoff. On `gpt-5.4-mini`, adaptive routing occupies the entire cost–accuracy Pareto frontier (80.7% at \$0.18/q vs. 78.1%/\$0.90 Ensemble and 76.8%/\$1.67 Predictor–Critic). The same Pareto ordering replicates on `gpt-5.4-nano`, where the Orchestrator is about 13× cheaper than the top Ensemble with no statistically significant accuracy gap; the Orchestrator's cost–quality advantage further extends to Claude Sonnet 4.5, DeepSeek v4 Flash, and Gemini 2.5 Flash, so the result is not a single-model artefact. A two-stage diagnostic explains the win as *selective spending*: the Orchestrator concentrates compute on questions where its cheap direct baseline is uncertain.

1. Introduction

Test-time compute scaling spends extra inference compute on a single query through search, deliberation, self-consistency, or multi-agent debate (Snell et al., 2024; Wei et al., 2022; Wang et al., 2023; Du et al., 2023; Lifshitz et al., 2025; Kim et al., 2026). Its strongest results come from verifiable domains (OpenAI et al., 2026; Guo et al., 2025): candidate solutions can be checked, so extra compute supports generate–verify–select. Forecasting future events has no such verifier, the outcome is unobserved at decision time, so additional compute can only improve the agent's information state, not certify outcomes. LLM forecasters are also typically retrieval-augmented, which raises two distinct evaluation risks (Halawi et al., 2024; Karger et al., 2025; Paleka et al., 2025; Dai et al., 2025): (i) *data contamination*, where retrieved pages leak the post-resolution outcome (a retrospective-backtesting risk), and (ii) *forecast outsourcing*, where the agent reads external probability sources (prediction-market odds, analyst quotes, expert forecasts) instead of reasoning from primary evidence. The first failure mode is specific to backtesting; the second remains relevant prospectively. Because we evaluate retrospectively on resolved questions, we must control for both, and we do so with a shared web-filter LLM (Section 2). We ask: *when no test-time verifier is available, how should inference compute be spent across questions, uniformly or selectively?*

We instantiate three test-time **compute-allocation policies**: static depth (*Predictor–Critic*), static breadth (*Ensemble*), and adaptive routing (*Hierarchical Orchestrator*). All three share base model, prompts, retrieval stack, metrics, and compute accounting; only the allocation rule varies. Full motivation, related work, and the policy formalism are in Sections A.1, A.2, F and G.

Contributions. (i) A contamination-controlled benchmark of 228 label-balanced post-cutoff ForecastBench questions, with a web-filter LLM whose metadata-based proxy recall on 847431 page decisions is 89.7%–92.1% (Sections A.3 and V). (ii) A controlled three-architecture sweep showing that adaptive routing occupies the cost–accuracy Pareto frontier on every base model in our sweep, with the same ordering replicated across `gpt-5.4-mini`, `gpt-5.4-nano`, Claude Sonnet 4.5, DeepSeek v4 Flash, and Gemini 2.5 Flash. (iii) A two-stage *selective-delegation* diagnostic: on three of five base models the Orchestrator's per-question LLM cost correlates with direct-baseline uncertainty, and direct-baseline uncertainty predicts paired Brier improvement on all five (Sections A.5 and H.0).

2. Setup

Benchmark and contamination control. We curate 228 binary ForecastBench questions (Karger et al., 2025) that all resolve strictly after the knowledge cutoff of every base model used here, are label-balanced (114 True / 114 False),

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

and preserve the topical mix of 12 ForecastBench partitions. All policies retrieve through a shared web-filter LLM that screens for the two contamination modes introduced in Section 1: post-cutoff timestamps and post-resolution coverage (*data contamination*), and any content that would let the agent outsource the forecast (*forecast outsourcing*: prediction-market odds, analyst probabilities, explicit outcome quotes). A metadata-based proxy audit estimates filter recall at 89.7%–92.1%. Curation, contamination protocol, and audit details are in Sections A.3, I and V.

Architectures. **Predictor–Critic** (T, E) refines a single predictor's forecast over E critic exchanges with predictor tool budget T, following Tree-of-Thoughts (Yao et al., 2023a) and Reflexion (Shinn et al., 2023). **Ensemble** (K, R) draws K independent predictors with R inter-agent revision rounds and combines them via an LLM aggregator, generalising self-consistency (Wang et al., 2023) and multi-agent debate (Du et al., 2023; Liang et al., 2024). **Hierarchical Orchestrator** (D, T) delegates up to D sub-tasks to sub-agents whose roles it defines at runtime, following LATS (Zhou et al., 2024) and AutoGen / MetaGPT (Wu et al., 2024a; Hong et al., 2024). Sub-agents share tool budget T. The full architecture descriptions, delegation modes, and ReAct-style compute-accounting upper bounds are in Sections A.3.1, C and E.

Metrics. We report accuracy Acc $= N^{-1}\sum_q \mathbb{1}[\hat{y}_q = y_q]$ (unparseable forecasts $\hat{y}_q = \perp$ counted incorrect) and Brier score (with $\tilde{p}_q = 0.5$ on $\perp$). For cost we report total USD (agent + web filter), agent-only cost, total tokens, LLM API calls, measured tool calls, and theoretical LLM calls. Statistical significance uses paired bootstrap (10k resamples) and exact McNemar tests (Sections A.3, S and T).

Configurations. The main `gpt-5.4-mini` sweep covers 22 configurations ($K \in \{1,4\}, R \in \{1,2,4\}$ for Ensemble; $T \in \{2,5\}, E \in \{0,1,2,5\}$ for Predictor–Critic; $D \in \{0,2,3,5\}, T \in \{1,2\}$ for Orchestrator). We replicate the full sweep on `gpt-5.4-nano` (right panel of Figure 2 and lower block of Table 1; full per-configuration results in Sections K and U) and run delegation-enabled Orchestrator scans on Claude Sonnet 4.5, DeepSeek v4 Flash, and Gemini 2.5 Flash. Architecture details are in Sections A.3 and E; the evaluation pipeline is summarised in Figure 1.

3. Results

Adaptive routing dominates the cost–quality frontier. On `gpt-5.4-mini` the Hierarchical Orchestrator occupies the entire observed cost–accuracy Pareto frontier (Figure 2; Table 1). Its best configuration (D=2, T=2, freeform) reaches 80.7% accuracy at \$0.175/question, against 78.1%/\$0.897 for the best Ensemble (K=1, R=2) and 76.8%/\$1.665 for the best Predictor–Critic (T=5, E=5):

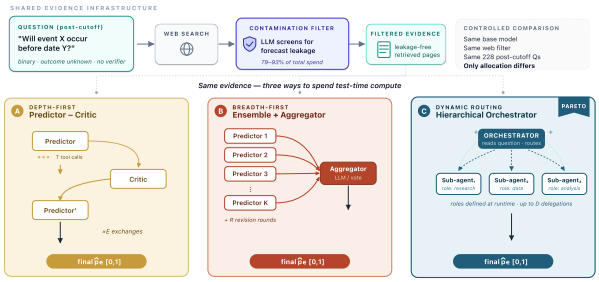


Figure 1. Evaluation pipeline and compute-allocation policies. A shared contamination-controlled retrieval interface feeds three policies: static depth (Predictor–Critic), static breadth (Ensemble + Aggregator), and adaptive routing (Hierarchical Orchestrator).

a 5.1× and 9.5× cost ratio respectively. The cost–quality result replicates on `gpt-5.4-nano`, where the best Orchestrator (D=3, T=2) is about 13× cheaper than the best Ensemble (K=4, R=4) at no statistically resolved accuracy gap (paired bootstrap 95% CI $[-4.4, +8.3]$ pp, McNemar exact p=0.68; right panel of Figure 2 and `nano` block of Table 1). The dominance is preserved under agent-only cost, LLM-call counts, and token counts (Section L). Raw Brier favours deeper Predictor–Critic configurations, but a single round of post-hoc Platt or temperature scaling closes most of the calibration gap and reduces best-per-architecture ECE by 34–42%, so the cost–quality ordering carries to the calibrated probabilistic outputs (Section O). At N=228 the peak accuracies sit within a 4.0-pp window and pairwise paired bootstrap CIs all straddle zero (McNemar $p \in \{0.25, 0.49, 0.75\}$); the conservative robust claim is therefore cost–quality dominance, not raw-accuracy supremacy (Section T). The full findings and the long-form Brier interpretation are in Section A.4.

Within-architecture scaling saturates. Inside every architecture, accuracy is non-monotone in budget (Figure 3). In Ensemble, going from K=1, R=1 to K=4, R=4 raises accuracy by 4.2 pp but multiplies cost 5.2×; in Predictor–Critic, moving from T=2, E=5 to T=5, E=5 adds substantial cost for only 0.5 pp; in the Orchestrator, jumping beyond D=2, T=2 leaves accuracy flat or worse at higher cost. Frontier gains are concentrated in a small number of low-budget transitions, especially for Predictor–Critic; the marginal-gain curves in Figure 3 all decay by roughly

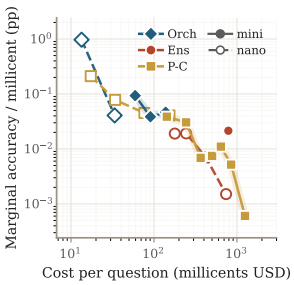


Figure 3. Marginal accuracy gain per millicent along each efficient frontier. Early budget transitions deliver almost all of the per-architecture accuracy; later compute saturates.

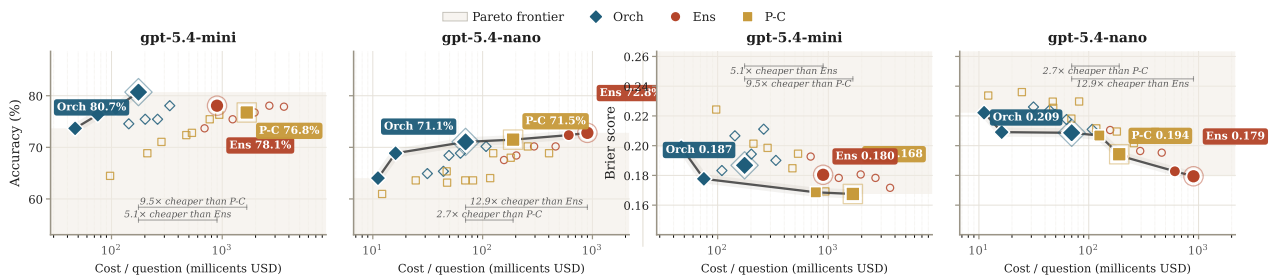


Figure 2. Cost–quality Pareto on `gpt-5.4-mini` (left) and `gpt-5.4-nano` (right). The Hierarchical Orchestrator occupies the low-cost accuracy frontier on both base models; Brier panels show the complementary calibration view. Cost is total pipeline (agent + web filter); agent-only and call-count proxies preserve the ordering (Section L).

Table 1. Best configuration per architecture on the 228-question benchmark for the `gpt-5.4-mini` (main sweep) and `gpt-5.4-nano` (replication) base models. Bold marks the best value within each model block. Full per-configuration tables are in x Section K.

Architecture	Best config	Acc. (%)	Brier	Cost ($/q)	Tokens/q (M)	LLM calls/q
`gpt-5.4-mini` (main sweep)						
Pred.–Critic	T=5, E=5	76.8	**0.168**	1.665	2.31	53.7
Ensemble	K=1, R=2	78.1	0.180	0.897	1.41	25.9
Orchestrator	D=2, T=2, freeform	**80.7**	0.187	**0.175**	**0.22**	**9.0**
`gpt-5.4-nano` (replication)						
Pred.–Critic	T=5, E=2	71.5	0.194	0.189	0.98	24.1
Ensemble	K=4, R=4	**72.8**	**0.179**	0.899	5.88	135.2
Orchestrator	D=3, T=2, freeform	71.1	0.209	**0.070**	**0.32**	**12.9**

an order of magnitude after the first one or two efficient points, so further compute on a fixed evidence pipeline buys diminishing accuracy regardless of how it is allocated. The implication is that "how to allocate" starts to dominate "how much to spend" beyond a small low-cost regime; the Orchestrator's selectivity is what lets it stay on the steep early portion of its own frontier (Section 4). Per-architecture heatmaps and saturation fits in Sections M and Q.

Mechanism: selective spend, not just less spend. The Orchestrator's architecture-level advantage comes from *selective macro-allocation* under a fixed compute ceiling, not from a uniformly smaller mandated budget. Table 2 contrasts theoretical and measured agent-side LLM calls per question. The best small Orchestrator (D=2, T=2) saturates its $(D{+}1) + D(T{+}1) = 9$-call ceiling; pushing to D=5, T=2 raises the ceiling to 21 but only 17.6 calls are used on average (median 18, tool calls 7.6/10), with no accuracy gain. Ensemble and Predictor–Critic, by contrast, instantiate their static schedules close to their theoretical maxima (108% and 81% respectively), so cost grows nearly mechanically with K, R, T, E. Adaptive routing therefore has a stopping mechanism the static policies lack. The aggregate gap is consistent with both a uniform under-spend (the agent stops below its ceiling on every question) and a selective under-spend (the agent stops earlier on easy questions and persists on harder ones); distinguishing the two requires conditioning per-question Orchestrator spend on question difficulty, which is what Section 4 does directly (per-configuration overlay in Sections A.4.3 and D).

Table 2. Theoretical vs. measured agent-side LLM calls per question on the `mini` sweep. Theoretical budgets exclude web-filter calls. Static schedules instantiate budgets close to their theoretical maxima; the Orchestrator saturates well below its ceiling at the larger delegation knob.

Architecture	Config.	Theory	Actual	Used
Ensemble	K=1, R=2	24.0	25.9	108%
Pred.–Critic	T=5, E=5	66.0	53.7	81%
Orchestrator	D=2, T=2	9.0	9.0	100%
Orchestrator	D=5, T=2	21.0	17.6	84%

Filter cost dominates total spend. The contamination-control web-filter LLM accounts for 83–95% of total spend across the `mini` sweep (median 91.8%; Table 3). The Orchestrator's lower share (83–93%, $0.15/q at its best configuration) follows from its smaller, history-conditioned retrieval volume; Ensemble and Predictor–Critic retrieve at fixed multiples of K and T and pay $0.85/q and $1.53/q at their top configurations, so the ranking is not a normalisation artefact. External evidence handling, not agent deliberation, is the systems bottleneck, and the conclusion is not specific to retrospective backtesting: only the post-resolution timestamp screen is backtesting-specific, while the forecast-

Table 3. Representative `mini` cost split. The web-filter LLM is the dominant share of total spend across all three architectures; the Orchestrator's lower share follows from its history-conditioned, smaller retrieval volume.

Configuration	Web share	Total $/q
Orchestrator $D{=}2, T{=}2$	87.9%	0.175
Orchestrator $D{=}5, T{=}2$	87.7%	0.336
Ensemble $K{=}1, R{=}2$	94.4%	0.897
Pred.–Critic $T{=}5, E{=}5$	91.9%	1.665

Table 4. Best delegation-enabled Orchestrator configuration per base model in the five-model scan ($D \in \{2, 3\}$, freeform). Accuracy spans 71.1–80.7%; cost spans nearly two orders of magnitude. Full per-configuration table in Section U.

Base model	Best config	Correct	Cost ($/q)
`gpt-5.4-mini`	$D{=}2, T{=}2$ freeform	184/228	0.175
Claude Sonnet 4.5	$D{=}2, T{=}2$ freeform	183/228	0.912
DeepSeek v4 Flash	$D{=}3, T{=}1$ freeform	183/228	0.037
Gemini 2.5 Flash	$D{=}3, T{=}2$ freeform	163/228	0.165
`gpt-5.4-nano`	$D{=}3, T{=}2$ freeform	162/228	0.070

outsourcing screen still inspects retrieved page content in prospective deployment, so deployment cost lies between the agent-only and total proxies. Full decompositions are in Section W.

Cross-model robustness. The cost–quality result is not an OpenAI-only artefact. Table 20 reports the best delegation-enabled Orchestrator per base model: `gpt-5.4-mini`, Claude Sonnet 4.5, and DeepSeek v4 Flash all reach 80.3–80.7% (183–184/228 correct) at costs that span two orders of magnitude ($0.037/q for DeepSeek to $0.912/q for Claude); `gpt-5.4-nano` and Gemini 2.5 Flash sit lower at 71.1–71.5% but at $0.07–$0.40/q, well below the best non-Orchestrator configuration on the same base. The best D is model-dependent ($D{=}2$ for mini and Claude; $D{=}3$ elsewhere), but on every model the Orchestrator's selected configuration sits at the low-cost end of its frontier, so the qualitative "adaptive routing dominates the cost–quality frontier" result replicates across all five base models we ran. Full per-configuration listings are in Section U.

4. Selective delegation diagnostic

Cost–quality dominance alone does not say whether the Orchestrator spends *less broadly* (uniform under-spend) or *less selectively* (concentrating compute on questions where extra compute is more likely to help). We test selectivity using direct-baseline uncertainty u_0 as a computable proxy for marginal compute value. Stage 1 asks whether per-question Orchestrator LLM cost correlates with u_0; Stage 2 asks whether u_0 correlates with realised paired Brier improvement of the Orchestrator over the direct baseline. Table 5 shows that Stage 1 is positive on three of five base models and Stage 2 is positive on all five, consistent with selective rather than uniform under-spending. The full setup, signal definitions, robustness checks, and Brier decomposition are in Sections A.5 and H.0.

Table 5. Two-stage selective-delegation diagnostic at $D{=}2, T{=}2$. Stage 1: $\rho_S(u_0, \text{cost})$, per-question orchestrator LLM cost vs. direct-baseline uncertainty (one-sided p). Stage 2: $\rho_S(u_0, \widehat{M}^{\text{Brier}})$, uncertainty vs. paired Brier improvement (two-sided p). Stage 1 is positive on $3/5$ models; Stage 2 on all five.

	Stage 1		Stage 2	
model	ρ	p (1-sided)	ρ	p (2-sided)
gpt-5.4-mini	$+0.145$	0.028	$+0.303$	< 0.001
gpt-5.4-nano	$+0.144$	0.030	$+0.220$	0.001
claude-sonnet-4.5	$+0.135$	0.042	$+0.178$	0.007
deepseek-v4-flash	-0.072	0.328	$+0.211$	0.004
gemini-2.5-flash	$+0.015$	0.826	$+0.207$	0.002

5. Discussion and Conclusion

In a verifier-free domain, the binding question is not how much inference compute to spend but *where*. Adaptive routing Pareto-dominates static breadth and depth on every base model in our sweep, and the mechanism is consistent with *selective* rather than uniform under-spending: the Orchestrator hits a tight agent-side ceiling on easy questions and concentrates additional compute on the harder ones, while Ensemble and Predictor–Critic spend their budgets mechanically across questions. Beyond a small budget, additional critic exchanges or parallel predictors operate on the same retrieved evidence and largely recombine the same signal, so accuracy plateaus at a retrieval-bounded ceiling rather than a deliberation-bounded one. The Brier picture refines, rather than overturns, this story: deeper Predictor–Critic runs are nominally better calibrated, but a single round of post-hoc Platt or temperature scaling closes most of the gap (App. O), so the cost–quality ordering carries over to the calibrated outputs. Furthermore, since the contamination-control web filter is 83–95% of total spend in our sweep, future gains are most likely to come from better *evidence infrastructure*, structured indices, distilled or cached filters, and domain-restricted crawls, rather than from adding agent deliberation; we view the filter recall audit (App. V) as a starting point and not a substitute for human-annotated validation. Architectures make complementary errors, and accuracy varies across difficulty bands and topics (App. N, P); these are scope conditions on the headline cost–quality claim, not exceptions to it. The full discussion, the related work, and the formal framework are in App. A.6, F, G; limitations in App. X. We release the benchmark, aligned architecture implementations, contamination-filter pipeline, and analysis code to support further work on cost-aware forecasting systems.

References

Aggarwal, P., Madaan, A., Yang, Y., and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12375–12396. Association for Computational Linguistics, December 2023. doi: 10.18653/v1/2023.emnlp-main.761. URL https://aclanthology.org/2023.emnlp-main.761/.

Brown, B., Juravsky, J., Ehrlich, R. S., Clark, R., Le, Q. V., Re, C., and Mirhoseini, A. Large language monkeys: Scaling inference compute with repeated sampling, 2025. URL https://openreview.net/forum?id=0xUEBQV54B.

Chandak, N., Goel, S., Prabhu, A., Hardt, M., and Geiping, J. Scaling open-ended reasoning to predict the future, 2026. URL https://openreview.net/forum?id=fkYrply0w3.

Dai, H., Teehan, R., and Ren, M. Are llms prescient? a continuous evaluation using daily news as the oracle, 2025. URL https://arxiv.org/abs/2411.08324.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Gema, A. P., Hägele, A., Chen, R., Arditi, A., Goldman-Wetzler, J., Fraser-Taliente, K., Sleight, H., Petrini, L., Michael, J., Alex, B., Minervini, P., Chen, Y., Benton, J., and Perez, E. Inverse scaling in test-time compute, 2025. URL https://arxiv.org/abs/2507.14417.

Graves, A. Adaptive computation time for recurrent neural networks, 2017. URL https://arxiv.org/abs/1603.08983.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24. Curran Associates Inc., 2024. ISBN 9798331314385.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., and Schmidhuber, J. Metagpt: Meta programming for a multi-agent collaborative framework, 2024. URL https://arxiv.org/abs/2308.00352.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. Large language models cannot self-correct reasoning yet. 2024. URL https://arxiv.org/abs/2310.01798.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lfPkGWXLLf.

Kim, J., Yang, W., Niu, K., Zhang, H., Zhu, Y., Helenowski, E., Silva, R., Chen, Z., Iyer, S., Zaheer, M., Fried, D., Hajishirzi, H., Arora, S., Synnaeve, G., Salakhutdinov, R., and Goyal, A. Scaling test-time compute for agentic coding, 2026. URL https://arxiv.org/abs/2604.16529.

Lahib, A. E., Xia, Y.-J., Li, Z., Wang, Y., and Pi, X. Temporal leakage in search-engine date-filtered web re-

trieval: A retrospective forecasting case study, 2026. URL https://arxiv.org/abs/2602.00758.

Lee, S.-W., Yang, S., Kwak, D., and Siegel, N. Y. Advancing event forecasting through massive training of large language models: Challenges, solutions, and broader impacts, 2025. URL https://arxiv.org/abs/2507.19477.

Li, J., Zhang, Q., Yu, Y., Fu, Q., and Ye, D. More agents is all you need, 2024. URL https://arxiv.org/abs/2402.05120.

Li, Z., Wang, Y., Lahib, A. E., Xia, Y.-J., and Pi, X. Simulated ignorance fails: A systematic study of llm behaviors on forecasting problems before model knowledge cutoff, 2026. URL https://arxiv.org/abs/2601.13717.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. Encouraging divergent thinking in large language models through multi-agent debate. pp. 17889–17904, November 2024. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Lifshitz, S., McIlraith, S. A., and Du, Y. Multi-agent verification: Scaling test-time compute with multiple verifiers. *arXiv preprint arXiv:2502.20379*, 2025.

Lin, K., Snell, C., Wang, Y., Packer, C., Wooders, S., Stoica, I., and Gonzalez, J. E. Sleep-time compute: Beyond inference scaling at test-time, 2025. URL https://arxiv.org/abs/2504.13171.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., and Clark, P. Self-refine: Iterative refinement with self-feedback. 2023. URL https://openreview.net/forum?id=S37hOerQLB.

Manvi, R., Singh, A., and Ermon, S. Adaptive inference-time compute: LLMs can predict if they can do better, even mid-generation, 2025. URL https://openreview.net/forum?id=7tOc6h8bea.

Murphy, A. H. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, 1973.

OpenAI, :, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., Kumar, A., Saraiva, A., Vallone, A., Duberstein, A., Kondrich, A., Mishchenko, A., Applebaum, A., Jiang, A., Nair, A., Zoph, B., Ghorbani, B., Zhang, B., Rossen, B., Sokolowsky, B., Barak, B., McGrew, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., McKinzie, B., Eastman, B., Lugaresi, C., Bassin, C., Hudson, C., Li, C. M., de Bourcy, C., Voss, C., Shen, C., Zhang, C., Koch, C., Orsinger, C., Hesse, C., Fischer, C., Chan, C., Roberts, D., Kappler, D., Levy, D., Selsam, D., Dohan, D., Farhi, D., Mely, D., Robinson, D., Tsipras, D., Li, D., Oprica, D., Freeman, E., Zhang, E., Wong, E., Proehl, E., Cheung, E., Mitchell, E., Wallace, E., Ritter, E., Mays, E., Wang, F., Such, F. P., Raso, F., Leoni, F., Tsimpourlas, F., Song, F., von Lohmann, F., Sulit, F., Salmon, G., Parascandolo, G., Chabot, G., Zhao, G., Brockman, G., Leclerc, G., Salman, H., Bao, H., Sheng, H., Andrin, H., Bagherinezhad, H., Ren, H., Lightman, H., Chung, H. W., Kivlichan, I., O'Connell, I., Osband, I., Gilaberte, I. C., Akkaya, I., Kostrikov, I., Sutskever, I., Kofman, I., Pachocki, J., Lennon, J., Wei, J., Harb, J., Twore, J., Feng, J., Yu, J., Weng, J., Tang, J., Yu, J., Candela, J. Q., Palermo, J., Parish, J., Heidecke, J., Hallman, J., Rizzo, J., Gordon, J., Uesato, J., Ward, J., Huizinga, J., Wang, J., Chen, K., Xiao, K., Singhal, K., Nguyen, K., Cobbe, K., Shi, K., Wood, K., Rimbach, K., Gu-Lemberg, K., Liu, K., Lu, K., Stone, K., Yu, K., Ahmad, L., Yang, L., Liu, L., Maksin, L., Ho, L., Fedus, L., Weng, L., Li, L., McCallum, L., Held, L., Kuhn, L., Kondraciuk, L., Kaiser, L., Metz, L., Boyd, M., Trebacz, M., Joglekar, M., Chen, M., Tintor, M., Meyer, M., Jones, M., Kaufer, M., Schwarzer, M., Shah, M., Yatbaz, M., Guan, M. Y., Xu, M., Yan, M., Glaese, M., Chen, M., Lampe, M., Malek, M., Wang, M., Fradin, M., McClay, M., Pavlov, M., Wang, M., Wang, M., Murati, M., Bavarian, M., Rohaninejad, M., McAleese, N., Chowdhury, N., Chowdhury, N., Ryder, N., Tezak, N., Brown, N., Nachum, O., Boiko, O., Murk, O., Watkins, O., Chao, P., Ashbourne, P., Izmailov, P., Zhokhov, P., Dias, R., Arora, R., Lin, R., Lopes, R. G., Gaon, R., Miyara, R., Leike, R., Hwang, R., Garg, R., Brown, R., James, R., Shu, R., Cheu, R., Greene, R., Jain, S., Altman, S., Toizer, S., Toyer, S., Miserendino, S., Agarwal, S., Hernandez, S., Baker, S., McKinney, S., Yan, S., Zhao, S., Hu, S., Santurkar, S., Chaudhuri, S. R., Zhang, S., Fu, S., Papay, S., Lin, S., Balaji, S., Sanjeev, S., Sidor, S., Broda, T., Clark, A., Wang, T., Gordon, T., Sanders, T., Patwardhan, T., Sottiaux, T., Degry, T., Dimson, T., Zheng, T., Garipov, T., Stasi, T., Bansal, T., Creech, T., Peterson, T., Eloundou, T., Qi, V., Kosaraju, V., Monaco, V., Pong, V., Fomenko, V., Zheng, W., Zhou, W., Zhan, W., McCabe, W., Zaremba, W., Dubois, Y., Lu, Y., Chen, Y., Cha, Y., Bai, Y., He, Y., Zhang, Y., Wang, Y., Shao, Z., and Li, Z. Openai o1 system card, 2026. URL https://arxiv.org/abs/2412.16720.

Paleka, D., Goel, S., Geiping, J., and Tramèr, F. Pitfalls in evaluating language model forecasters. 2025. URL

https://arxiv.org/abs/2506.00723.

Qin, J. and Andriushchenko, M. Quantsightbench: Evaluating llm quantitative forecasting with prediction intervals, 2026. URL https://arxiv.org/abs/2604.15859.

Schaeffer, R., Kazdan, J., Hughes, J., Juravsky, J., Price, S., Lynch, A., Jones, E., Kirk, R., Mirhoseini, A., and Koyejo, S. How do large language monkeys get their power (laws)? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=QqVZ28qems.

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

Schoenegger, P., Tuminauskaite, I., Park, P. S., and Tetlock, P. E. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy. 2024. URL https://arxiv.org/abs/2402.19379.

Schuster, T., Fisch, A., Gupta, J., Dehghani, M., Bahri, D., Tran, V. Q., Tay, Y., and Metzler, D. Confident adaptive language modeling. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=uLYc4L3C81A.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Curran Associates Inc., 2023.

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Stechly, K., Valmeekam, K., and Kambhampati, S. On the self-verification limitations of large language models on reasoning and planning tasks. 2024. URL https://arxiv.org/abs/2402.08115.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024a. URL https://openreview.net/forum?id=BAakY1hNKS.

Wu, Y., Sun, Z., Li, S., Welleck, S., and Yang, Y. Scaling inference computation: Compute-optimal inference for problem-solving with language models. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*, 2024b. URL https://openreview.net/forum?id=j7DZWSc8qu.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. R. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=5Xc1ecxO1h.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models. 2023b. URL https://arxiv.org/abs/2210.03629.

Ye, C., Hu, Z., Deng, Y., Huang, Z., Ma, M. D., Zhu, Y., and Wang, W. Mirai: Evaluating llm agents for event forecasting, 2024. URL https://arxiv.org/abs/2407.01231.

Zhou, A., Yan, K., Shlapentokh-Rothman, M., Wang, H., and Wang, Y.-X. Language agent tree search unifies reasoning acting and planning in language models, 2024. URL https://openreview.net/forum?id=6LNTSrJjBe.

Zilberstein, S. Using anytime algorithms in intelligent systems. *AI Magazine*, 17(3):73–83, 1996.

Appendix overview

The appendix is organised in two parts. **Part A** reproduces, in full, the sections that were compressed in the 4-page main body, so that all narrative material from the original long-form draft remains available without modification. The headline figures and tables that appear in the main body are not duplicated here; cross-references to Figures 1 and 2 and Tables 1 and 5 in this appendix resolve to the main-body floats. **Part B** contains the original supporting appendices unchanged from the long-form draft: compute-accounting derivations, the formal compute-allocation framework, full tables and per-configuration results, robustness checks, calibration recovery, the dataset card, the cross-model and cost-breakdown breakdowns, and limitations.

A. Extended main-body sections (full long-form versions)

This appendix part reproduces, in full, the six main-body sections that were compressed in the 4-page workshop body. Section labels match those in the original long-form draft so that cross-references from the main body resolve here.

A.1. Extended Introduction (full version)

This appendix section reproduces the full version of the Introduction. The compressed introduction in the main body cites this appendix for background, motivation, and the long-form contribution list.

Test-time compute scaling, which spends more inference-time computation on a single query through search, deliberation, self-consistency, or multi-agent debate, has emerged as a powerful complement to model scaling (Snell et al., 2024; Wei et al., 2022; Wang et al., 2023; Du et al., 2023; Lifshitz et al., 2025; Kim et al., 2026). Its most convincing successes, however, come from domains with test-time *verifiability* (OpenAI et al., 2026; Guo et al., 2025). In mathematics, code, and many structured reasoning tasks, candidate solutions can often be checked before an answer is returned: programs can be executed, derivations can be inspected for consistency, and independent reasoning paths can be compared against known constraints. Additional compute therefore supports a search-and-selection process in which incorrect candidates can be generated, detected, and discarded.

Forecasting provides a qualitatively different setting for test-time compute scaling. LLM forecasting systems are increasingly being used to predict unresolved future events and time-indexed quantities from incomplete evidence. This literature is broad and includes a variety of approaches: retrieval-augmented forecasters for open-ended future questions (Halawi et al., 2024; Chandak et al., 2026), ensemble and crowd-aggregation approaches (Schoenegger & Park, 2023; Schoenegger et al., 2024), dynamic forecasting across diverse real-world domains (Karger et al., 2025), tool-using agents for temporal event prediction (Ye et al., 2024), and quantitative benchmarks for calibrated, non-binary forecasts (Qin & Andriushchenko, 2026).

The key difference from mathematics and code is that the target outcome is unobserved at decision time. A forecasting agent may retrieve more evidence, generate competing hypotheses, elicit independent forecasts, critique assumptions, or aggregate multiple estimates; however, it cannot observe a correctness signal for any candidate forecast before the event resolves. Additional inference-time compute therefore cannot instantiate the same generate–verify–select loop that succeeds in verifier-rich domains. Instead, it can only improve the agent's information state or reduce decision errors conditional on that information, while the realised outcome remains subject to irreducible epistemic uncertainty. This makes it unclear whether test-time compute gains observed on verifiable reasoning benchmarks should transfer to forecasting.

When no test-time verifier is available, what does additional inference compute buy, and how should it be spent across questions, uniformly or selectively on the ones where it pays off?

This question is especially delicate because LLM forecasting is usually retrieval-augmented: agents gather external evidence before predicting (Halawi et al., 2024; Karger et al., 2025). In retrospective backtests, retrieval can leak the resolved outcome through post-resolution pages; even prospectively, agents can outsource the forecast by reading prediction markets or analyst probability quotes rather than reasoning from primary evidence (Paleka et al., 2025; Dai et al., 2025). We therefore evaluate all models under a shared web-filtering stack that targets outcome leakage in retrospective backtests and, at inference time, can exclude market odds, analyst probabilities, or other disallowed sources that would bias the model's own forecast (Section A.3) (Paleka et al., 2025).

We study test-time scaling in forecasting through the lens of compute allocation. We instantiate three test-time **compute-allocation policies**: decision rules that distribute inference-time compute across web-based evidence gathering, independent

predictors, critique, and delegation (see Section G.1). We evaluate all policies under a single contamination-controlled pipeline, which we make publicly available together with the benchmark, model outputs, and analysis code. Our design isolates the effect of compute allocation by holding fixed the base model, web-filtering layer, cutoff date, metrics, and compute-accounting rules across all reported runs. We make the following contributions.

[leftmargin=*, nosep]

1. **A framework for evaluating compute allocation in verifier-free forecasting.** We curate a balanced post-cutoff benchmark of 228 binary ForecastBench questions and implement a web-evidence filtering layer that screens retrieved pages for forecast leakage and post-resolution contamination in the multi-agent setting; a metadata-based proxy audit estimates that it blocks roughly 90–92% of likely leakage signals. On this benchmark we evaluate three aligned compute-allocation policies: static depth (Predictor–Critic), static breadth (Ensemble + Aggregator), and adaptive routing (Hierarchical Orchestrator), under the same questions, tools, prompts, contamination safeguards, and compute accounting. The complete sweep is run with both `gpt-5.4-nano` and `gpt-5.4-mini`, so shared findings can be separated from model-specific capability or formatting effects.
2. **Demonstrate that adaptive routing dominates the cost–quality frontier, and the systems bottleneck is retrieval, not reasoning.** The Hierarchical Orchestrator occupies the cost–accuracy Pareto frontier on every base model in our sweep, matching or beating both static-depth (Predictor–Critic) and static-breadth (Ensemble) architectures at substantially lower cost. The ordering replicates across two complete model sweeps (`gpt-5.4-mini`, `gpt-5.4-nano`) and across a five-model Orchestrator scan, ruling out a single-model artefact. Across configurations, the contamination-controlled web filter dominates total spend, identifying retrieval-side compute, not agent deliberation, as the systems bottleneck. The cost–quality ordering carries over to probabilistic outputs once calibration is normalised post-hoc. Full numbers, statistical-significance tests, and per-configuration tables in Section A.4.
3. **A mechanistic explanation of the cost–quality dominance.** We propose the *selective-delegation hypothesis*, in which adaptive routing wins because it concentrates compute on questions where extra compute is more likely to help, and test it via a two-stage diagnostic. We find that on three of five base models, the Orchestrator's per-question LLM compute correlates with direct-baseline uncertainty (a computable proxy for marginal compute value), and that direct-baseline uncertainty correlates with Brier improvement across all five models.

A.2. Extended framing: compute allocation in verifier-free forecasting

Full version of the framing section, including the formal statement of the four hypotheses H1–H4, the evidence-limited ceiling intuition, and the connection to the appendix-only theoretical framework.

As established in the introduction (Section A.1), forecasting is a verifier-free setting: at inference time the binary resolution $Y \in \{0, 1\}$ is unobserved, so candidate forecasts cannot be checked before they are returned. We therefore treat test-time forecasting as a *compute-allocation problem*: a policy π distributes inference-time actions, including web/tool calls, predictor calls, critique, aggregation, and sub-agent delegations, across the trajectory of answering question q under budget C, and emits a forecast at termination. We compare three such policies: static depth (Predictor–Critic), static breadth (Ensemble + Aggregator), and history-dependent adaptive routing (Hierarchical Orchestrator). The full sequential formalism is in Appendix G.1.

One conceptual idea anchors the framing. Once a policy has gathered its evidence, no amount of additional reasoning on that fixed evidence can push accuracy past what is in principle decidable from that evidence, there is an *evidence-limited ceiling*. Compute that *changes* the evidence (more retrieval, broader queries) can lower the ceiling; compute that only adds reasoning passes on fixed evidence cannot. Appendix G.2, G.4 give the formal Bayes-risk decomposition $L(\pi, C) = L^*(R_\pi(C)) + \Delta(\pi, R_\pi(C))$ that makes this precise: L^* is the Bayes risk under the policy's evidence, Δ is the decision suboptimality on top of it.

This framing yields four empirical predictions that we test:

[leftmargin=*, nosep]

- **(H1) Adaptive routing can improve the cost–quality frontier** (Section A.4.1).
- **(H2) Policies should show diminishing returns toward similar evidence-limited ceilings** under shared retrieval and contamination filtering (Section A.4.2).

- **(H3) Retrieval-side compute can dominate total cost** when contamination control is implemented as an evidence-filtering layer (Section A.4.4).
- **(H4) Selective-delegation hypothesis.** The cost–quality advantage of adaptive routing can be mechanistically explained by *selective spending* – concentrating compute on the questions where extra compute is more likely to help (easy questions terminate early, ambiguous questions spend more), not just spending less on every question (Section A.5).

A.3. Extended methods: benchmark, metrics, architectures

Full version of the methods section, including the benchmark curation procedure, the four contamination safeguards, the metric definitions, the architecture descriptions, and the configuration sweep.

Benchmark: 228 balanced post-cutoff questions. Our benchmark is 228 binary forecasting questions subsampled from the ForecastBench pool (Karger et al., 2025) under three joint constraints: (i) every question's `market_info_close_datetime` falls strictly after the knowledge cutoff of every base model used here (`gpt-5.4-mini`, `gpt-5.4-nano`, Claude Sonnet 4.5, DeepSeek v4 Flash, Gemini 2.5 Flash), with resolutions in February 2026–January 2027; (ii) exactly 114 True / 114 False, removing the majority-class shortcut; (iii) topical mix preserved across 12 ForecastBench partitions (Politics/Geopolitics, Technology, Sports, Economics/Finance, and others). Full curation procedure, per-partition counts, topical breakdown, and the dataset card are in Appendix I.

Contamination safeguards. We address the four pitfalls catalogued by Paleka et al. (2025): **(i) Temporal leakage**, by using a *web-filter LLM* to inspect every retrieved page and remove any page containing timestamps or event references after the question-specific cutoff date, which we derive from the ForecastBench market metadata (the midpoint before `market_info_close_datetime`); **(ii) Forecast contamination**, by removing pages containing explicit probability estimates, prediction-market odds, analyst forecasts, or price targets; **(iii) Backtesting artefacts**, by using `market_info_midpoint_datetime`, rather than the later `close_datetime`, so agents forecast from an information state before the market has closed; and **(iv) Skewed base rates**, by curating a 50/50 True/False set (Section A.3).

As an audit of this filter, we run a metadata-based proxy recall check over 847431 page-level filter decisions from the nano and mini runs. Pages whose metadata indicates a likely leakage signal are blocked with proxy recall 89.7% (Wilson 95% CI $[89.5, 89.8]\%$), rising to 92.1% ($[91.9, 92.3]\%$) under a stricter metadata definition. This is not a substitute for human annotation; metadata can both over- and under-identify true contamination, but it provides a scalable sanity check that the filter catches the large majority of obvious leakage candidates. Appendix V gives the full audit breakdown and candidate-miss analysis.

Evaluation metrics and compute accounting. For a benchmark of N questions $\{q_1, \ldots, q_N\}$, let $y_q \in \{0, 1\}$ denote the realised resolution of question q, and let policy π emit a probabilistic forecast $\hat{p}_q \in [0, 1]$ together with a parsed binary answer $\hat{y}_q \in \{0, 1\} \cup \{\perp\}$, where $\perp$ indicates an unparseable output.

Forecast quality. Our headline decision metric is strict accuracy, $\mathrm{Acc}(\pi) = N^{-1} \sum_q \mathbb{1}[\hat{y}_q = y_q]$, with $\hat{y}_q = \perp$ counted as incorrect. We also report Brier score, $\mathrm{Brier}(\pi) = N^{-1} \sum_q (\tilde{p}_q - y_q)^2$, setting $\tilde{p}_q = 0.5$ when $\hat{y}_q = \perp$, so that probabilistic quality is not hidden by thresholding. Conditional versions of both metrics are computed after dropping $\perp$, always paired with the answer rate $\#\{q : \hat{y}_q \neq \perp\}/N$.

Cost and uncertainty. We report seven compute proxies per configuration: total USD cost, agent-only LLM cost, total tokens, prompt tokens, LLM API calls, measured tool calls, and theoretical LLM calls (Section A.3.1). We compare paired binary outcomes with an exact McNemar test and report bootstrap (10 000-resample) 95% confidence intervals for accuracy differences across architectures.

A.3.1. Architectures

We compare three multi-agent compute-allocation policies. All three take the same question, call the same web-search tool through the same contamination filter, share the same base model, and output a single probability $\hat{p} \in [0, 1]$; they differ only in *how* they spend additional inference-time compute. **(A) Predictor–Critic** (depth-first) refines a single predictor's forecast over E critic exchanges with predictor tool budget T, in the spirit of Tree-of-Thoughts (Yao et al., 2023a) and Reflexion (Shinn et al., 2023). **(B) Ensemble + Aggregator** (breadth-first) draws K independent predictors with R optional inter-agent revision rounds and combines them via LLM aggregator or confidence-weighted vote, generalising self-consistency (Wang et al., 2023) and multi-agent debate (Du et al., 2023; Liang et al., 2024). **(C) Hierarchical Orchestrator** (dynamic routing) delegates up to D sub-tasks to sub-agents with tool budget T whose roles it defines at runtime, following LATS-style

adaptive search (Zhou et al., 2024) and AutoGen/MetaGPT-style orchestration (Wu et al., 2024a; Hong et al., 2024). The five integer knobs (T, E, K, R, D) are the only architectural degrees of freedom; everything else, including base model, prompts, filter, tool suite, and metric implementation, is shared, and we sweep them to trace each architecture's efficient frontier (Section A.4). Full per-architecture descriptions, delegation modes, and design rationale are in Appendix E.

Compute accounting Worst-case per-question LLM- and tool-call upper bounds for each architecture, derived from the shared ReAct-style loop (Yao et al., 2023b), are deferred to Appendix C and used to compute utilisation $\mathrm{actual}/\mathrm{theoretical}$ in Section A.4.3.

A.3.2. Experimental Setup

The main experiment is a controlled OpenAI `gpt-5.4-mini` sweep over the three compute-allocation policies in Figure 1. Ensemble varies $K \in \{1, 4\}$ and $R \in \{1, 2, 4\}$; Predictor–Critic varies $T \in \{2, 5\}$ and $E \in \{0, 1, 2, 5\}$; and the Orchestrator varies $D \in \{0, 2, 3, 5\}$ and $T \in \{1, 2\}$, using direct tool access at $D{=}0$ and freeform delegation for $D \geq 2$. The main analysis reports 22 `mini` configurations: 6 Ensemble, 8 Predictor–Critic, and 8 Orchestrator.We use OpenAI `gpt-5.4-nano` as a full-sweep replication in the appendix, and report an additional Orchestrator-only cross-model check over complete delegation-enabled configurations from Claude Sonnet 4.5, DeepSeek v4 Flash, Gemini 2.5 Flash, `gpt-5.4-mini`, and `gpt-5.4-nano` in Section A.4.5. All runs use the same benchmark, tools, web-filtering layer, metrics, and compute accounting. Further setup details, including baselines, variance estimation, and implementation notes, are in Appendix S.

A.4. Extended empirical findings: allocation, saturation, and the filter bottleneck

Full version of the empirical findings section. The main body summarises the headline cost–quality result, the theoretical-vs-actual utilisation gap, the filter-cost bottleneck, and the cross-model robustness check; this appendix gives the per-finding prose, the Brier interpretation, the within-architecture diminishing-returns analysis, and the statistical caveats. Figures and tables referenced below appear in the main body.

We use the contamination-controlled benchmark and compute accounting of Sections A.3 and A.3.2 to evaluate the three compute-allocation policies of Section A.2: static depth (Predictor–Critic), static breadth (Ensemble + Aggregator), and adaptive routing (Hierarchical Orchestrator). The main analysis is the controlled `gpt-5.4-mini` sweep over 22 configurations (6 Ensemble, 8 Predictor–Critic, 8 Orchestrator). All three policies share the same questions, tool stack, contamination filter, and cost accounting; only the allocation rule varies.

Cost convention. Unless otherwise stated, cost refers to *total pipeline cost*: agent/model calls plus the web-filter calls required for contamination-controlled backtesting. We report agent-only cost, token, and call-count proxies separately in the appendix; the qualitative Pareto ordering is preserved under all three (Appendix L). This distinction matters because part of the web-filter cost is specific to backtesting, especially the temporal-cutoff component, while the forecast-outsourcing filter remains relevant for prospective deployment. Retaining that filter still requires inspecting retrieved page content, so removing only the temporal backtesting check would not reduce the cost to the agent-only proxy. Prospective deployment costs would therefore lie between the total and agent-only proxies, and may remain closer to total pipeline cost when forecast-outsourcing filtering is retained.

Section A.2 motivates three qualitative hypotheses about what we expect to observe when these policies run head-to-head, and the experiments are designed to test each in turn: (**H1**) adaptive routing can occupy the cost–quality frontier because it can condition compute on the realised history; (**H2**) under a shared retrieval-and-filter stack, all policies are expected to saturate as their budgets grow; and (**H3**) when contamination control is implemented as page-level filtering, retrieval-side filtering may dominate the cost accounting. We also flag a statistical caveat at the outset: at $N{=}228$ and an inter-architecture peak-accuracy spread of ≤ 4.0 pp, raw-accuracy gaps among the three best configurations are not statistically resolved (Appendix T). The robust empirical claim of this section is therefore *cost–quality dominance of adaptive routing*, not raw-accuracy supremacy.

The section is organised around four findings followed by a robustness check. Section A.4.1 establishes the headline cost–quality result (Figure 2, Table 1); Section A.4.2 shows that within each architecture additional compute saturates and is sometimes non-monotone (Figure 3); Section A.4.3 identifies the underlying mechanism via a utilisation gap (Table 2); Section A.4.4 locates where the systems cost actually goes (Table 3); and Section A.4.5 verifies that the headline result is not a single-model artefact (Figure 4). Compute-proxy robustness (token- and call-based Paretos), calibration recovery, an

oracle architecture router, descriptive saturation fits, and the full `gpt-5.4-nano` sweep are reported in the appendix and surfaced in the Discussion (Section A.6) where they bear on the narrative.

A.4.1. Adaptive routing dominates the cost–quality frontier

Finding (C1, H1). The Hierarchical Orchestrator occupies the entire observed cost–accuracy Pareto frontier. Its best configuration is also the cheapest of the three best-per-architecture points, by a factor of $5.1\times$ relative to Ensemble and $9.5\times$ relative to Predictor–Critic.

The best Orchestrator configuration ($D{=}2, T{=}2$, freeform) reaches 80.7% accuracy at \$0.175/question. The best Ensemble configuration ($K{=}1, R{=}2$) reaches 78.1% at \$0.897/question, and the best Predictor–Critic configuration ($T{=}5, E{=}5$) reaches 76.8% at \$1.665/question. The Orchestrator is therefore the raw accuracy maximum of the `mini` sweep *and* the cheapest of the three best-per-architecture points. This conclusion does not depend on web-filter pricing: the same Orchestrator-only frontier appears under agent-only cost, LLM-call count, and token-count proxies (Appendix L). The Orchestrator's best configuration uses \$0.021 agent-only cost and 9.0 agent-side LLM calls per question, against the best Ensemble's \$0.050 and 25.9 calls. Under total-token accounting (agent plus web filter), the corresponding counts are 219k vs. 1.41M tokens.

The three peak accuracies sit within a 4.0-percentage-point window. Paired bootstrap 95% CIs for the pairwise accuracy gaps all straddle zero, and exact-McNemar tests give $p \in \{0.25, 0.49, 0.75\}$ (Appendix T). We therefore make the deliberately conservative headline claim that the Orchestrator is the strongest observed *cost–quality* point, not the statistically resolved peak in raw accuracy. The cost columns of Table 1 carry the headline.

Brier interpretation. The Brier column of Table 1 and the right panel of Figure 2 are best read as a *complementary calibration view* of the same configurations rather than a refinement of the accuracy ranking. Predictor–Critic's depth-first deliberation produces the best raw Brier in the `mini` sweep (0.168 at $T{=}5, E{=}5$ vs. 0.180 Ensemble and 0.187 Orchestrator), at roughly $9\times$ higher cost than the Orchestrator best. Under the Murphy decomposition (Section G.2, Eq. 2), most of this Brier gap is the *Reliability* (calibration) component, not the *Resolution* (discriminative) component. Empirical expected calibration error (ECE) summarises just the Reliability term, and a single round of post-hoc Platt or temperature scaling on a small held-out set reduces the best-per-architecture ECE by 34–42%, from raw ECE 0.13–0.14 down to 0.08–0.09 on held-out data (Appendix O, Table 16). The accuracy and Brier panels of Figure 2 therefore answer two different questions: "which architecture is most accurate?" and "which architecture is best calibrated before any post-hoc scaling?". The calibration gap of the cheap-Orchestrator regime closes after this scaling step, so the cost–quality ordering of the accuracy panel transfers to the calibrated probabilistic outputs.

If adaptive routing dominates at the architecture level, a natural next question is whether *within* each architecture additional compute keeps buying accuracy. We turn to that next.

A.4.2. Within-architecture scaling: diminishing and non-monotone returns

Finding (C2, H2). Inside every architecture, accuracy is non-monotone in budget: adding more predictors, exchanges, or delegations beyond a small sweet spot does *not* reliably improve performance. Frontier gains are concentrated in a small number of low-budget transitions, especially for Predictor–Critic (Figure 3).

The clearest examples of non-monotonicity are: in Ensemble, going from $K{=}1, R{=}1$ to $K{=}4, R{=}4$ raises accuracy by 4.2 percentage points but multiplies cost by $5.2\times$; in Predictor–Critic, moving from $T{=}2, E{=}5$ to $T{=}5, E{=}5$ adds substantial cost for only 0.5 percentage points; in Orchestrator, jumping beyond $D{=}2, T{=}2$ leaves accuracy flat or worse at higher cost. The full per-architecture heatmaps are in Appendix M; descriptive saturation fits to each architecture's efficient frontier and their robustness to the fitting procedure are reported in Appendix Q (Section Q.3) and Appendix R.

This is the systems half of the answer to "why doesn't more compute buy more accuracy?". The complementary question is what the Orchestrator does *differently* that lets it sit at the cheapest point on the joint frontier.

A.4.3. Actual vs. theoretical compute budgets

Finding (H1 mechanism). The Orchestrator's advantage comes from selective macro-allocation: it reaches the best cost–quality point with a small useful delegation budget. Larger Orchestrator budgets are only partially used and do not improve accuracy. Static architectures, by contrast, instantiate their predictor, critic, or ensemble schedules much more

mechanically across questions. Table 2 summarises the main utilisation contrast; the full per-configuration overlay is in Appendix D (Figure 5).

Concretely (Table 2), the best Orchestrator configuration ($D{=}2, T{=}2$) uses a compact agent-side budget: its measured LLM calls are 9.0 per question, matching the theoretical maximum $(D{+}1) + D(T{+}1) = 9$. This is far smaller than the best Ensemble (25.9 calls) and Predictor–Critic (53.7 calls). Increasing the Orchestrator delegation budget does not translate into proportional useful compute or higher accuracy. At $D{=}5, T{=}2$, the theoretical maximum is 21 agent LLM calls, but the empirical mean is 17.6 and the median is 18; measured tool calls average 7.6 against a theoretical maximum of 10. Ensemble and Predictor–Critic, by contrast, more directly instantiate their static macro-schedules: increasing predictors, rounds, tools, or critic exchanges increases cost mechanically even when accuracy does not improve. This makes H1 concrete: the Orchestrator wins by finding a smaller useful macro-budget, not by spending more inference compute.

Macro under-utilization in the Orchestrator alone cannot distinguish two mechanisms: uniform budget management, in which the Orchestrator spends less broadly, and selective allocation, in which it concentrates compute on questions where extra compute is more likely to help. We test this selectivity mechanism using direct-baseline uncertainty in Section A.5.

This mechanism accounts for the agent-side call budget, but not for the total-cost gaps in Table 1. If agent-side model usage is not the dominant source of spend, where does the rest of the cost go?

A.4.4. The retrieval-and-filter bottleneck

Finding (C3, H3). The web-filtering LLM, needed to prevent retrieval of post-resolution contamination or forecast-leaking content, dominates total cost across every configuration in the sweep. Agent reasoning is a minority share of total spend.

The web filter accounts for 83–95% of total spend across the `mini` sweep (median 91.8%). External-evidence handling, not agent deliberation, is the systems bottleneck. The Orchestrator's lower share (83–93%) follows directly from its smaller and history-conditioned retrieval volume; Ensemble and Predictor–Critic, which retrieve at fixed multiples of K and T respectively, sit near the upper end of the range (about 90–95%). This is evidence for H3: when contamination control is implemented as filtering on retrieved pages, retrieval-side compute dominates the accounting. The absolute decomposition in Appendix W shows that this pattern is not only a percentage-share effect: static breadth and depth also incur much larger total dollar costs. We discuss implications for future contamination-aware retrieval in Section A.6. The full percentage and dollar decompositions are in Appendix W (Figures 19 and 20). The remaining question is whether the cost–quality dominance result holds outside the `gpt-5.4` model family.

A.4.5. Cross-model robustness

Finding. The cost–quality dominance of adaptive routing is not a single-model artefact. A delegation-enabled Orchestrator comparison over the selected D settings confirms that orchestration carries the result across base-model families, while the best delegation depth is model-dependent. We scan Orchestrator runs from Claude Sonnet 4.5, DeepSeek v4 Flash, Gemini 2.5 Flash, `gpt-5.4-mini`, and `gpt-5.4-nano`. For this cross-model comparison, we focus on the selected delegation-enabled settings.

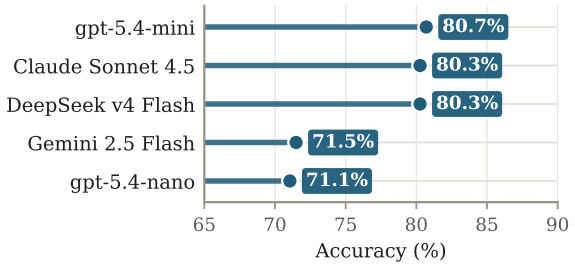


Figure 4. Best delegation-enabled Orchestrator per model ($D \in \{2, 3\}$). Per-model configuration, correct count, and cost are listed in Appendix U.

Two observations follow from Figure 4; the numeric table is in Appendix U. First, delegated orchestration is not an OpenAI-only artefact: `gpt-5.4-mini` reaches $184/228$ correct, while Claude Sonnet 4.5 and DeepSeek v4 Flash each reach $183/228$. Second, among the delegation-enabled Orchestrator settings we compare across models, the best configuration is not uniform: `mini` and Claude peak at $D{=}2, T{=}2$; DeepSeek peaks at $D{=}3, T{=}1$; Gemini and `nano` peak at $D{=}3, T{=}2$.

The complete `gpt-5.4-nano` architecture sweep is consistent with the same systems lesson. Its top Ensemble is slightly higher in raw accuracy than its top Orchestrator (72.8% vs. 71.1%), but the paired difference is not statistically resolved (bootstrap 95% CI $[-4.4, +8.3]$ pp, McNemar exact $p = 0.68$), and the Orchestrator is about $13\times$ cheaper ($0.070/question vs. $0.899/question; Appendix U). The robust nano conclusion

is similar accuracy at substantially lower cost, not a statistically decisive raw-accuracy ordering.

A.5. Extended selective delegation: routing on direct-baseline uncertainty

Full version of the selective-delegation section. The main body summarises the two-stage diagnostic and reports Table 5; this appendix gives the longer prose around the result.

To test whether the Orchestrator's low cost reflects selective routing rather than uniform under-spending, we use direct-baseline uncertainty u_0 as a computable proxy for where extra compute may help. The diagnostic has two stages: first, whether per-question Orchestrator spend correlates with u_0; second, whether u_0 correlates with realised paired Brier improvement. Table 5 shows that Stage 1 is positive on three of five models, while Stage 2 is positive on all five. This pattern is consistent with selective delegation, but remains correlational rather than causal; the full setup, signal definition, and robustness checks are in Appendix G.6.

A.6. Extended discussion and conclusion

Full version of the discussion and conclusion. The main body conclusion is a four-line summary; the long-form discussion of allocation, retrieval bottleneck, scope conditions, and contamination-aware retrieval directions is reproduced here.

We presented a controlled study of test-time compute for LLM forecasting on 228 balanced post-cutoff questions. In this verifier-free domain, the binding question is not simply how much inference compute to spend, but where to spend it. Across the `mini` sweep, adaptive routing is the best observed cost–quality allocation: the top Orchestrator reaches 80.7% accuracy at $\$0.175$/question, while the best static breadth and depth systems cost $5.1\times$ and $9.5\times$ more. The `nano` replication shows the same fast-saturation systems pattern even when its best raw-accuracy point is an Ensemble, and the cross-model Orchestrator snapshot suggests that delegated routing is not specific to one OpenAI model family.

The mechanism is selective allocation. Static Ensemble and Predictor–Critic schedules spend their budgets on every question, whereas the Orchestrator can stop early or delegate only when additional evidence appears useful. This explains both the low-cost frontier and the diminishing returns within each architecture. Once the filtered evidence has been incorporated, extra critic exchanges or additional parallel predictors often restate the same information rather than raising the evidence-limited ceiling. The Brier results refine, rather than overturn, this story: deeper Predictor–Critic runs are better calibrated before post-hoc scaling, but simple Platt or temperature scaling recovers much of the calibration gap (Appendix O).

The main systems bottleneck is evidence handling. The web filter accounts for 83–95% of total spend in the `mini` sweep, so future gains are likely to come from more efficient contamination-aware retrieval, such as structured indices, distilled filters, or domain-restricted crawls, rather than from simply adding agent deliberation. The metadata audit suggests that the filter catches most obvious leakage signals, but human annotation remains necessary before treating this as validated recall.

Several details are therefore best read as scope conditions. The architectures make complementary errors, as shown by the oracle-router and disagreement analyses in Appendix N; accuracy varies across difficulty bands and topics (Appendix P); and the pre-cutoff side experiment is only a diagnostic because training-time exposure may inflate absolute accuracy (Appendix J). The forecasting lesson is nevertheless consistent across these checks: when no verifier is available, compute allocation and evidence infrastructure matter more than raw compute volume. We release the benchmark, aligned architecture implementations, and analysis pipeline to support further work on cost-aware forecasting systems.

B. Societal impact

This work studies evaluation and compute allocation for forecasting agents, rather than deploying a forecasting system in a real decision setting. Its main positive impact is methodological: better contamination-controlled evaluations can make forecasting agents more auditable, cost-aware, and reproducible. If used responsibly, such tools could support domains where calibrated forecasts and transparent evidence use are valuable, such as policy analysis, scientific planning, public-health preparedness, financial risk monitoring, and resource allocation under uncertainty.

The same capabilities also create risks. More capable forecasting agents could be over-trusted in high-stakes settings, used to justify decisions without appropriate human review, or optimized toward narrow measurable objectives while missing social, legal, or ethical context. Forecasting systems may also amplify biases in their evidence sources, leak or overuse market and analyst forecasts, or create unequal access if useful forecasting tools are available only to well-resourced actors.

In adversarial settings, improved forecasts could support harmful strategic planning, market manipulation, or information operations.

Our contribution is therefore best viewed as an evaluation framework and not as a recommendation to automate consequential decisions. We mitigate some evaluation-specific risks by using post-cutoff questions, reporting uncertainty and limitations, filtering retrieved evidence for leakage, and releasing anonymized code and benchmark materials for audit. These measures do not address all deployment risks; real-world use would require domain-specific validation, human oversight, access controls where appropriate, and monitoring for misuse and distributional harms.

C. Compute-accounting derivation

The main text uses the compact upper bounds in Table 6. This appendix gives the derivation. Every tool-using agent implements the same ReAct-style loop: with tool budget B, it emits at most B tool calls and $B + 1$ LLM calls, where the extra LLM call is the final synthesis after the last observation. If the agent emits a final answer early, actual usage is lower; the bounds below are worst-case accounting ceilings.

Let M denote the fixed per-loop tool budget for agents whose budget is not swept. In the reported runs, $M{=}5$ for Ensemble predictors, the Ensemble aggregator, and the Predictor–Critic critic. The swept variable T is used for the Predictor–Critic predictor and the Orchestrator's delegated sub-agents.

Derivation. Each architecture composes ReAct loops in a different pattern. Counting loops, then summing $(B{+}1)$ LLM calls and B tool calls per loop with the appropriate per-loop budget $B \in \{T, M\}$, gives:

- **Ensemble.** K predictors each run the loop in each of $1{+}R$ rounds (initial plus R revisions), then a single aggregator runs one ReAct loop. All predictors and the aggregator use budget M, so

$$\mathrm{LLM}_{\text{theo}} = K(1{+}R)(M{+}1) + (M{+}1), \qquad \mathrm{Tool}_{\text{theo}} = K(1{+}R)M + M.$$

- **Predictor–Critic.** The predictor runs $1{+}E$ ReAct loops (initial plus E revisions), each with swept budget T; the critic runs E ReAct loops, each with fixed budget M. Summing predictor and critic contributions gives

$$\mathrm{LLM}_{\text{theo}} = (1{+}E)(T{+}1) + E(M{+}1), \qquad \mathrm{Tool}_{\text{theo}} = T(1{+}E) + ME.$$

- **Hierarchical Orchestrator.** For $D \geq 1$, the orchestrator itself acts as a router: its available high-level actions are sub-agent delegations, exposed through the same function-calling interface as ordinary tools. It emits at most $D{+}1$ LLM calls: one initial decision, one after each of up to D sub-agent returns, and a final synthesis call. Each delegated sub-agent runs a ReAct loop with swept budget T, contributing up to $T{+}1$ LLM calls and T tool calls per delegation. Thus

$$\mathrm{LLM}_{\text{theo}} = (D{+}1) + D(T{+}1), \qquad \mathrm{Tool}_{\text{theo}} = DT,$$

  understood as the worst case in which all D delegations are used and every sub-agent saturates its tool budget. At $D{=}0$, the orchestrator collapses to a single direct tool-using ReAct loop with budget T, so $\mathrm{LLM}_{\text{theo}}^{D=0} = T{+}1$ and $\mathrm{Tool}_{\text{theo}}^{D=0} = T$.

Table 6. Theoretical per-question compute upper bounds used for utilisation analysis. These are accounting bounds, not typical-use predictions; measured cost is reported separately in Section A.3.

Architecture	Swept knobs	LLM calls upper bound	Tool calls upper bound
Ensemble	K, R	$K(1{+}R)(M{+}1) + (M{+}1)$	$K(1{+}R)M + M$
Predictor–Critic	T, E	$(1{+}E)(T{+}1) + E(M{+}1)$	$T(1{+}E) + ME$
Orchestrator, $D{=}0$	T	$T{+}1$	T
Orchestrator, $D \geq 1$	D, T	$(D{+}1) + D(T{+}1)$	DT

We use these formulas to compute utilisation rates $\mathrm{actual}/\mathrm{theoretical}$ in Section A.4.3. The gap between these worst-case ceilings and observed usage is a mechanism: static schedules instantiate their macro budget on every question, whereas the Orchestrator can stop before consuming all delegations.

D. Full theoretical-vs-actual utilisation overlay

Table 2 in the main text reports representative configurations. Figure 5 shows the full per-configuration overlay for the `gpt-5.4-mini` sweep.

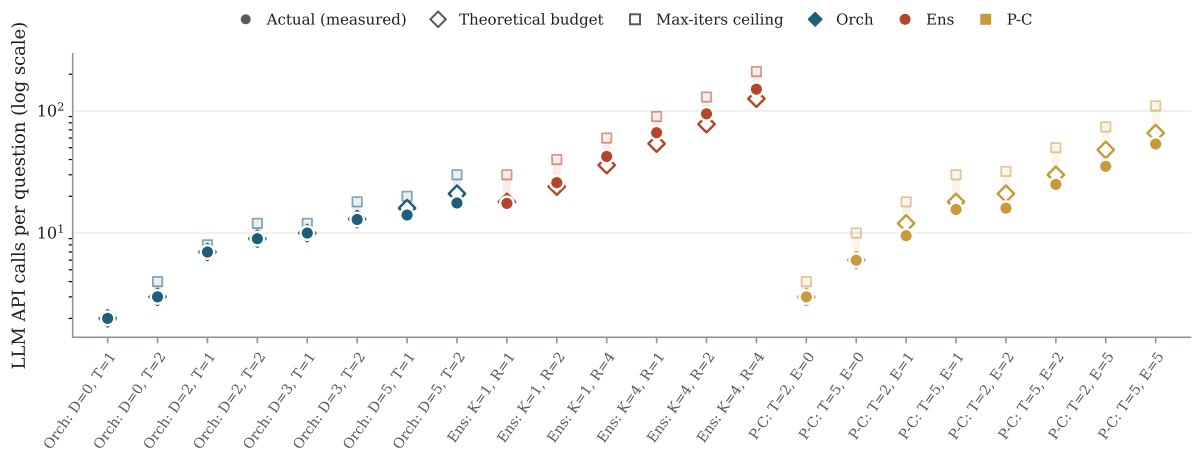


Figure 5. Theoretical vs. actual LLM calls per question on the `gpt-5.4-mini` sweep. Filled discs are measured calls, open diamonds are theoretical per-config budgets, and faded squares are loop max-iters ceilings.

E. Architectures: full descriptions

This appendix gives the full per-architecture descriptions condensed in Section A.3.1. The three multi-agent compute-allocation policies all take the same question, call the same web-search tool through the same contamination filter, share the same base model, and output a single probability $\hat{p} \in [0, 1]$. They differ only in *how* they spend additional inference-time compute.

A. Predictor–Critic (depth-first). A single predictor with T tool calls produces an initial forecast; a critic challenges its reasoning and the predictor revises over E exchanges, yielding $1 + E$ rounds total. Structurally this instantiates the depth-first deliberate-search style of Tree-of-Thoughts (Yao et al., 2023a) and Reflexion-style self-critique (Shinn et al., 2023), restricted to a fixed-width linear search to keep the budget interpretable.

B. Ensemble + Aggregator (breadth-first). K independent predictors each generate an initial forecast (round 0), optionally exchange information for R further revision rounds (rounds 1 through R, for $1+R$ cycles total per predictor), then are combined by either an LLM aggregator or a confidence-weighted majority vote. This generalises self-consistency (Wang et al., 2023) with an explicit inter-agent communication phase along the lines of multi-agent debate (Du et al., 2023; Liang et al., 2024).

C. Hierarchical Orchestrator (dynamic routing). A central orchestrator analyses each question and dynamically delegates up to D sub-tasks to sub-agents whose roles it defines at runtime based on the question's content and on the output of each previous sub-agent. Each sub-agent receives a tool budget T and has access to the full tool suite (web search, financial APIs, reasoning); the orchestrator is forced to produce a final answer once D delegations have been used. We use the *freeform* delegation mode throughout: at $D{=}0$ the orchestrator predicts directly with tool access (no sub-agents); for $D \geq 2$ the orchestrator is free to define each sub-agent's objective from scratch. A separate *predefined* mode (fixed Research / Data / Analysis sub-agents) was implemented but is not part of the reported sweep (Section X). The adaptive-routing structure follows the LATS family of agentic search architectures (Zhou et al., 2024) and the orchestration abstraction of AutoGen / MetaGPT (Wu et al., 2024a; Hong et al., 2024), specialised here to forecasting and constrained to a hard delegation budget so that the policy class is comparable to (A) and (B).

Compute knobs. The five integer knobs (T, E, K, R, D) are the only architectural degrees of freedom; everything else, including base model, prompts, filter, tool suite, and metric implementation, is shared. We sweep them to trace each architecture's efficient frontier (Section A.4). At $D{=}0$ the Orchestrator predicts directly without delegating. The

implementation choice of *freeform* vs. *predefined* sub-agent roles, used only in side experiments, is documented in Section X.

F. Related Work

We organise prior work into four blocks that align with our three contributions: *(i) LLM forecasting systems and benchmarks* and *(iv) temporal contamination and evaluation validity* position our contamination-controlled evaluation protocol (C1); *(iii) multi-agent and deliberative inference* positions our aligned three-architecture comparison (C2); and *(ii) test-time compute and inference-time scaling* positions our empirical Pareto analysis and descriptive saturation fits (C3).

LLM forecasting systems and benchmarks. Halawi et al. (2024) establish that retrieval-augmented LLMs can approach human forecasters with a fixed inference pipeline (retrieve → summarise → predict → aggregate); we treat this as the natural starting point for our study. Schoenegger & Park (2023); Schoenegger et al. (2024) study LLM ensembles in tournaments, and Karger et al. (2025) provide the dynamic benchmark from which our 228-question dataset is curated. Chandak et al. (2026) take a training-time approach: they synthesise forecasting questions from daily news and fine-tune Qwen3 on the resulting *OpenForesight* dataset using reinforcement learning, showing that their 8B *OpenForecaster* matches much larger proprietary models on held-out future events; Lee et al. (2025) argue similarly that massive forecasting-specific training is now within reach and outline the noisiness, knowledge-cutoff, and reward-design challenges that scaling such training would have to overcome. This line of work is complementary to ours: it scales *training* compute with a single-pass inference step, while we hold the model weights fixed and study how different ways of allocating *test-time* compute affect the same base model. Qin & Andriushchenko (2026) concurrently introduce *QuantSightBench*, evaluating LLM *quantitative* forecasting via prediction-interval coverage on continuous quantities, and report systematic overconfidence in every frontier model they test (`gemini-3.1-pro`, `grok-4`, `gpt-5.4`, all $\geq$ 10 percentage points short of the 90% coverage target). Relative to this line of work, our paper asks the next question, *once the forecasting pipeline is in place, what does extra test-time compute buy us?*, and we study three structurally distinct ways of spending that compute under a unified budget framework.

Test-time compute and inference-time scaling. Snell et al. (2024) show test-time compute can be more effective than parameter scaling on math; OpenAI et al. (2026) and Guo et al. (2025) demonstrate that large-scale reinforcement learning with verifiable rewards turns inference-time deliberation into substantial gains on math and code; Wu et al. (2024b); Brown et al. (2025); Li et al. (2024) fit inference-time scaling laws for math and code, and report sub-logarithmic returns from repeated sampling, with Schaeffer et al. (2025) giving a recent theoretical account of the observed power-law shape. Closest in spirit to our work, Kim et al. (2026) scale test-time compute for *agentic* coding by representing prior rollouts as compact summaries and selecting/refining among them; their setting still sits in the verifier-rich coding regime (SWE-Bench, Terminal-Bench), whereas our forecasting domain has no test-time verifier at all. Chain-of-thought (Wei et al., 2022) and self-consistency (Wang et al., 2023) are the canonical breadth-style references, and Lifshitz et al. (2025) demonstrates verifier-based search. Adaptive-computation policies, including learnt halting (Graves, 2017; Schuster et al., 2022), anytime algorithms (Zilberstein, 1996), adaptive self-consistency (Aggarwal et al., 2023), and self-aware mid-generation routing (Manvi et al., 2025), formalise the intuition that history-dependent allocation strictly contains static schedules, but have not been analysed under a Pareto framework with a corresponding theoretical bound. Lin et al. (2025) broaden the picture by showing that some useful compute can be moved out of test time entirely; this is a useful contrast point because forecasting limits the applicability of sleep-time compute: future-question content is by definition not available in advance. Crucially, Gema et al. (2025) document *inverse* test-time scaling, where additional reasoning compute *degrades* performance on certain tasks; this is directly relevant to our diminishing-returns finding (Section A.4.2). Our positioning relative to this line of work is empirical rather than theoretical: we report the first *controlled three-architecture Pareto study* of test-time compute in a verifier-free domain (Section A.4), with descriptive saturating-compute fits to each architecture's efficient frontier (Section G.5). The architecture ranking is *motivated*, not predicted, by a policy-containment argument (Theorem G.1), since whether any specific implemented orchestrator realises the containment advantage on a particular benchmark is itself empirical. The dominant spend in our cost analysis is the contamination-filter LLM, not the agent, a systems observation we discuss in (iv) below.

Multi-agent and deliberative inference. Multi-agent debate (Du et al., 2023; Liang et al., 2024) and self-refinement (Shinn et al., 2023; Madaan et al., 2023) explore sequential argumentation and revision. Tree-of-Thoughts (Yao et al., 2023a) formalises depth-first deliberate search over intermediate reasoning states, structurally close to our Predictor–Critic architecture. ReAct (Yao et al., 2023b) interleaves reasoning with tool calls and is the basis of our agent action loop; LATS (Zhou et al., 2024) unifies ReAct-style tool use with ToT-style search and adds external feedback, structurally close to

our Hierarchical Orchestrator. AutoGen (Wu et al., 2024a) and MetaGPT (Hong et al., 2024) are general orchestration frameworks; we specialise their adaptive-policy idea to forecasting and tie it to a formal containment argument. Self-correction without an external signal has been shown to be unreliable (Huang et al., 2024; Stechly et al., 2024), which motivates our use of an external critic and a delegating orchestrator rather than purely introspective refinement. We instantiate these strands as three *aligned* architectures so that depth, breadth, and dynamic routing can be compared under a single budget framework.

Temporal contamination and evaluation validity. Paleka et al. (2025) catalogue evaluation pitfalls, including temporal leakage, retrieval contamination by prediction-market prices, skewed base rates, which we explicitly mitigate. Li et al. (2026) show that simply prompting an LLM to "ignore" post-cutoff information is unreliable, motivating our requirement that all evaluation questions resolve strictly *after* the model's knowledge cutoff (Section A.3). Lahib et al. (2026) demonstrate that even date-filtered web search leaks post-resolution contamination evidence, substantially inflating apparent forecasting accuracy; this is the direct empirical motivation for the dedicated contamination-filter LLM in our pipeline (Section A.3), and the implementation choice that makes filter-side compute the dominant share of total spend in our cost analysis. Dai et al. (2025) provide a complementary view from the model side, showing that LLM forecasting accuracy degrades systematically as pre-training data ages and that retrieval-augmented prompting only partially recovers the gap, consistent with our observation that, once contamination is controlled, additional reasoning compute on a fixed (filtered) evidence set saturates quickly. Together, these recent results support our position that contamination control is a first-class component of any test-time-compute claim about forecasting.

G. A compute-allocation framework for test-time forecasting

The empirical results in Section A.4 show three patterns that, taken together, are not adequately explained by existing test-time-compute scaling theory (Snell et al., 2024; Wu et al., 2024b; Brown et al., 2025): (i) a dynamic-routing policy Pareto-dominates static breadth and static depth on every compute proxy; (ii) all three architectures saturate near a similar accuracy ceiling; (iii) marginal accuracy per extra unit of compute drops by an order of magnitude after one or two points on each architecture's efficient frontier. We give a small decision-theoretic framework that provides qualitative intuition for these three observations. The framework combines standard ingredients, namely policy-class containment and the Bayes risk of decisions under fixed evidence, with a descriptive empirical curve fit on the 228-question Pareto data.

Notation reference. The framework uses the standard random-variable convention: *uppercase* symbols denote random variables, *lowercase* symbols denote their realisations, and a hat $(\hat{\cdot})$ marks a quantity emitted by the policy (an estimate). Empirical metrics in Section A.3 are Monte-Carlo estimates of the population quantities below.

Table 7. Notation used throughout Section G and the appendices. Empirical analogs (with sample subscripts) are defined in Section A.3.

Symbol	Type	Meaning
Q	random variable	A forecasting question
$Y \in \{0, 1\}$	random variable	The (unobserved at test time) resolution of Q
q, y_q	realisations	A specific question and its realised resolution
$\hat{y} \in \{0, 1\}$	policy output (RV)	Binary forecast emitted by π on Q
$\hat{p} \in [0, 1]$	policy output (RV)	Probabilistic forecast emitted by π on Q
$\hat{y}_q$, $\hat{p}_q$	realisations	Policy outputs on a specific question q
$\hat{p}_{\text{cal}}$	post-hoc calibrated	Output of Platt / temperature scaling on $\hat{p}$
π	function	Test-time-compute policy (architecture + config)
$C \geq 0$	scalar	Compute budget (USD per question, or token count)
$\mathcal{A}$	set	Action set: {tool call, sub-agent delegation, predictor invocation, critic exchange}
h_t	sequence	History of evidence + intermediate predictions at step t
$R(C)$	set	Evidence retrieved by π within budget C
$L(\pi, C)$	population scalar	Expected zero–one loss of π at budget C
L_{ev}, L_{rs}, L_{cal}	population scalars	Three components of L (Equation (1))
α_{arch}	population scalar	Saturation ceiling of architecture's scaling-law fit
β_{arch}	population scalar	Compute-recoverable headroom
τ_{arch}	population scalar	Compute time-constant (smaller $\Rightarrow$ faster saturation)

G.1. Compute-allocation policies: formal definitions

Notation convention. We follow the standard random-variable convention: *uppercase* Q, Y denote random variables (a forecasting question and its unobserved-at-test-time resolution); *lowercase* q, y_q denote realisations of those variables on a specific question; and the policy's outputs on that question are $\hat{p}_q \in [0,1]$ (probabilistic forecast) and $\hat{y}_q \in \{0,1\}$ (parsed binary answer). The empirical metrics in Section A.3 are Monte-Carlo estimates of the population quantities defined here.

Let Q be a forecasting question and $Y \in \{0,1\}$ its (unobserved at test time) resolution. A *test-time-compute policy* is a budget-constrained sequential decision rule. Let $\mathcal{A}$ denote the action set: {tool call, sub-agent delegation, predictor invocation, critic exchange, aggregation step}. Starting from $h_0 = (Q)$ and budget $b_0 = C$, the policy chooses actions according to a kernel

$$\pi_t(\cdot \mid h_t, b_t) \in \Delta(\mathcal{A} \cup \{\text{TERMINATE}\}),$$

where $h_t = (Q, a_{<t}, o_{<t})$ contains all previous actions and observations, and b_t is the remaining budget. If $a_t \in \mathcal{A}$, the evaluation runner executes the corresponding model, tool, or sub-agent operation, receives observation o_t, pays the measured cost, and updates both h_t and b_t. If $a_t = \text{TERMINATE}$, or if no feasible action remains under the budget, the policy stops at time τ and a terminal map g_π emits

$$(\hat{p}, \hat{y}) = g_\pi(h_\tau), \qquad \hat{p} \in [0,1], \quad \hat{y} \in \{0,1\}.$$

The evidence object $R_\pi(C)$ is the projection of h_τ onto the retrieved pages, filtered tool outputs, intermediate forecasts, and other observations available when the policy stops. Deterministic policies or implementations are included as the degenerate case where each $\pi_t(\cdot \mid h_t, b_t)$ is a point mass.

Every architecture in this paper instantiates such a policy:

- *Predictor–Critic* $\pi_{\text{PC}}^{T,E}$ runs T tool calls, then alternates predictor and critic agents for E exchanges according to a fixed macro schedule.
- *Ensemble* $\pi_{\text{Ens}}^{K,R}$ runs K predictors in parallel and exchanges information for R rounds according to a fixed macro schedule.
- *Hierarchical Orchestrator* $\pi_{\text{Orch}}^{D,T}$ dynamically defines and delegates up to D sub-tasks, with both the *assignment* of each sub-task and the *role* of the sub-agent chosen *conditional* on h_t (the question text and the previous sub-agents' return values). At $D{=}0$ the orchestrator acts directly with tool access.

We call a policy *static at the macro-allocation level* if, conditional on the question Q and step t, its next high-level action distribution does not depend on the realised history h_t, and *adaptive* otherwise:

$$\Pi_{\text{static}} = \{\pi : P_\pi(a_t \mid Q, t, h_t) = P_\pi(a_t \mid Q, t)\},$$

$$\Pi_{\text{adapt}} = \{\pi : P_\pi(a_t \mid Q, t, h_t) \text{ may depend on } h_t\}.$$

Thus $\Pi_{\text{static}} \subsetneq \Pi_{\text{adapt}}$: a static macro-allocation policy is a special case of an adaptive policy that ignores h_t, while there exist adaptive policies whose high-level actions genuinely depend on the realised history.

Static at the level of macro compute allocation. In our implementation π_{PC} and π_{Ens} are static at the level of *macro compute allocation*: the number of predictor calls, critic exchanges, and ensemble rounds is fixed in advance by the parameters (T, E) and (K, R), and is not modulated by intermediate outputs. Within each individual ReAct invocation, however, the agent does choose its next tool query in response to prior observations, so PC and Ensemble are not *strictly* static at the fine-grained tool-action level. The relevant distinction for our framework and for the policy-class containment argument in Section G.3 is at the macro-allocation level: π_{Orch} is the only policy in our sweep whose macro schedule (which sub-agent is delegated to next, with what role, and whether to delegate at all) depends on h_t.

G.2. Two error decompositions: accuracy and calibration

We organise the analysis around two complementary error decompositions, one for thresholded accuracy and one for the probabilistic forecast.

Accuracy decomposition. Let $L(\pi, C) = \Pr(\hat{y} \neq Y)$ be the population zero–one error of policy π at compute budget C. Each policy induces an evidence set $R_\pi(C)$, the pages, queries, and tool outputs π retrieves under budget C, and the loss splits cleanly into two non-negative pieces:

$$L(\pi, C) \;=\; \underbrace{L^*(R_\pi(C))}_{\text{Bayes risk under } \pi\text{'s evidence}} \;+\; \underbrace{\Delta(\pi, R_\pi(C))}_{\text{decision sub-optimality}}, \tag{1}$$

where $L^*(R_\pi(C))$ is the Bayes risk given the evidence distribution induced by π (Section G.4), and $\Delta(\pi, R_\pi(C)) \geq 0$ is the additional error from π failing to make the Bayes-optimal call *given that specific evidence*. Increasing reasoning compute on a fixed evidence set can only reduce Δ, never L^*. Note that $R_\pi(C)$ is in general policy-dependent: different architectures may issue different queries at the same budget, so L^* is itself architecture-conditional rather than a single shared quantity. We argue in Section G.4 that under our shared retrieval-and-filter stack the gap between the policies' $R_\pi(C)$ distributions is small, which is why their fitted ceilings end up close (Section Q.3).

Calibration as a separate axis. Calibration error is *not* an additive component of the zero–one loss in (1): a perfectly-calibrated forecaster can be at chance accuracy and a miscalibrated forecaster can have identical thresholded decisions to a well-calibrated one. Calibration naturally decomposes proper scoring rules. We use the standard Murphy decomposition (Murphy, 1973) of the Brier score,

$$\text{Brier}(\pi) \;=\; \underbrace{\text{Reliability}(\pi)}_{\text{calibration error}} \;-\; \underbrace{\text{Resolution}(\pi)}_{\text{discriminative power}} \;+\; \underbrace{\text{Uncertainty}(Y)}_{\text{base-rate variance}}, \tag{2}$$

in which the Reliability term is the only component the calibration recovery procedure (Section O) targets: it is the squared gap between the predicted probability and the observed frequency in each bin, weighted by bin size, and is approximated by the empirical expected calibration error (ECE) we report. ECE is therefore an estimator of the Reliability term alone, not of the full Brier score. Resolution measures how strongly the forecaster separates events with different empirical frequencies, and Uncertainty depends only on the base rate of Y (it is the same for every forecaster on the same benchmark).

The two decompositions are independent and we analyse them separately throughout the paper: accuracy via (1) and the Pareto frontier in Section A.4, calibration via (2) and ECE in Section O. Combining them into a single composite objective would require a custom loss; we do not.

G.3. Allocation dominance

Proposition G.1 (Adaptive policies weakly dominate static ones)**.** *For any compute budget $C \geq 0$, $\min_{\pi \in \Pi_{adapt}} L(\pi, C) \leq \min_{\pi \in \Pi_{static}} L(\pi, C)$, with strict inequality whenever there exists a positive-probability event in which the optimal action depends on h_t.*

Proof sketch. Containment: any static π_s is also in Π_{adapt}, so the adaptive optimum is at least as good. For strictness, observe that an adaptive policy can re-allocate compute when h_t reveals a sub-task is unproductive (e.g. skip a Data sub-agent when the question is purely qualitative); the static policy must spend the same compute regardless. A standard policy-improvement argument (Graves, 2017; Schuster et al., 2022; Aggarwal et al., 2023; Manvi et al., 2025) converts any positive-probability detection event into a strict inequality. □

Proposition G.1 is a *necessary* but not sufficient condition for our empirical headline. It establishes that the optimal adaptive policy can in principle match any static schedule on any compute budget; it does *not* guarantee that any specific implemented orchestrator dominates any specific implemented ensemble or predictor–critic on any particular benchmark, since implementation quality of all three architectures will determine how close each is to the corresponding minimum. Whether our concrete Hierarchical Orchestrator realises the containment advantage on the 228-question benchmark is therefore an *empirical* question, which we test in Section A.4.

G.4. Decision-theoretic evidence ceiling

The *shape* of the saturation pattern is bounded by a standard decision-theoretic argument. Let $\eta(R) = \Pr(Y = 1 \mid R)$ denote the posterior probability of $Y = 1$ given evidence set R. For a binary classification loss, the Bayes-optimal decision error given R is

$$L^*(R) \;=\; \mathbb{E}_R\big[\, \min\big(\eta(R),\, 1 - \eta(R)\big) \,\big], \tag{3}$$

and any decision rule π acting on R satisfies $\Pr(\hat{y} \neq Y) \geq L^*(R)$. In our setting the evidence set depends on the budget, $R = R(C)$, so the smallest population error *achievable by any policy operating on the same evidence* is $L_{\text{ev}}(C) \geq L^*(R(C))$. Two consequences follow:

- **Similar ceilings under shared evidence infrastructure.** Architectures that share the question set, the retrieval backend, and the contamination filter operate on approximately the same evidence distribution and are therefore bounded by approximately the same $L^*(R(C))$. Different policies may issue different queries and accumulate different evidence sets in any given run, so we expect ceilings to be *similar in expectation but not identical*, consistent with the empirical ≤ 0.04 pairwise gap between the best α values per architecture (Table 1).
- **Reasoning compute on fixed evidence is bounded by Bayes risk.** Doubling E in PC or R in Ensemble at fixed $R(C)$ cannot push accuracy past $1 - L^*(R(C))$. Only compute that changes the evidence (more retrieval, broader queries) can lower the floor.

Why verifier-free is qualitatively different. The split itself is generic to classification; the forecasting-specific point is what test-time compute can do to its two terms. In verifier-rich domains, a verifier produces an additional observation V about candidate correctness (for example, program execution or a symbolic validity check), so the effective evidence becomes $R^+_\pi(C) = (R_\pi(C), V)$. Conditioning on a more informative state cannot increase the Bayes-optimal error in expectation, so the verifier raises the evidence-limited ceiling: $L^*(R^+_\pi(C)) \leq L^*(R_\pi(C))$. Verifier feedback can also reduce Δ by making the implemented selection rule closer to Bayes-optimal on the evidence it has. In forecasting, the realised outcome Y is unavailable at test time and there is no analogous correctness observation for a candidate forecast: extra deliberation on fixed evidence can only target Δ, never L^*. This is the formal sense in which verifier-free reasoning is structurally different from verifier-rich reasoning, and it motivates our focus on *allocation*: which actions raise the evidence-limited ceiling vs. which only target Δ, rather than on aggregate inference budget.

Cost-breakdown as a methodological observation. A separate, empirical observation is that in our implementation the contamination filter accounts for 83–95% of total spend in the `mini` sweep (Figure 20). We do not derive this share from theory because it reflects two implementation choices, namely the use of an LLM-based filter (rather than a regex / smaller-model classifier) and the use of a relatively cheap base model whose own per-call cost is small compared to the per-page filter cost. The figure therefore quantifies *the cost of running a contamination-controlled forecasting evaluation* under our specific stack, and identifies the filter as a clear systems lever for future cost reduction; it is not an information-theoretic prediction.

G.5. A descriptive saturation fit for the efficient frontier

The Bayes-risk bound in (3) suggests that, on a fixed retrieval-and-filter stack, accuracy as a function of compute should saturate. We fit a three-parameter exponential saturation *descriptively* to each architecture's efficient frontier:

$$\text{Acc}_{\text{arch}}(C) \;=\; \alpha_{\text{arch}} \;-\; \beta_{\text{arch}}\, \exp(-C/\tau_{\text{arch}})\,, \tag{4}$$

where α_{arch} is the asymptotic ceiling, β_{arch} the compute-recoverable headroom, and τ_{arch} the architecture-specific compute time constant.

We emphasise that this is a *descriptive* fit, not a scaling "law" in the strict sense: each architecture's efficient frontier contributes only 3–5 non-dominated points, which is too few to estimate a three-parameter family with tight confidence (Appendix Q reports bootstrap CIs that are correspondingly wide). We therefore use (4) as a compact summary of the cost–accuracy curves rather than as a predictive scaling model. The framework supplies two qualitative predictions and one explicit non-prediction:

1. *Approximately shared ceilings.* $\alpha_{\text{Orch}}, \alpha_{\text{Ens}}, \alpha_{\text{PC}}$ should be *approximately* equal because the three architectures share the same question set, retrieval backend, and contamination filter, so they operate on similar (in expectation) evidence distributions and therefore similar Bayes risks $L^*(R(C))$ (Section G.4). The framework does not predict exact equality, since different policies may issue different queries and accumulate different evidence sets in any given run; we therefore expect the pairwise gap $\max_{i,j} |\alpha_i - \alpha_j|$ to be small but non-zero.

2. *Faster saturation for adaptive routing.* The history-adaptive policy class *can* reach a target accuracy at no greater compute than either static class (Proposition G.1). If this advantage is realised by our specific implementation, the orchestrator's τ should be smaller than either static τ. The proposition does not by itself guarantee this for our concrete implementation; that is an empirical question.

3. *Non-prediction: order between the two static policies.* The framework does *not* order τ_{Ens} and τ_{PC}. Their relative speed depends on per-step costs and on which dimension, width K or depth E, offers more diversity per unit cost in a particular regime. We report the empirical ordering in Section Q.3 as data, not as a confirmed prediction.

Empirical fits on the 228-question data are reported in Section Q.3 and Appendix Q, with confidence intervals from a non-parametric bootstrap.

G.6. What this framework explains, and what it does not

The framework offers *qualitative* intuitions for: (a) why adaptive routing has a structural opportunity to dominate (§G.3; whether our concrete implementation realises this opportunity is empirical), (b) why all three architectures saturate near approximately the same ceiling under shared evidence infrastructure (§G.4), (c) why diminishing returns are sharp once the evidence ceiling is approached (§G.5), and (d) why calibration error is best analysed on a different axis than thresholded accuracy (§G.2, (2)). It does *not* predict the absolute value of α (which depends on the retrieval pipeline and the model's prior knowledge), nor the exact ratio of saturation time-constants (which we estimate descriptively, not predictively).

Cross-model hypothesis: delegation budgets and base-model strength interact. The framework extends to a cross-model comparison under one mild assumption: when only the base LLM changes, the per-question evidence ceiling $L_{\text{ev}}(C)$ is unchanged (same web filter, same retrieval), and only the reasoning sub-optimality term $L_{\text{rs}}(\pi, C)$ shifts: weaker models incur a larger L_{rs} at every C. Because the orchestrator's policy class is strictly larger than either static policy's (Theorem G.1), it has more freedom to compensate for a larger L_{rs} by re-routing compute. Two qualitative hypotheses follow:

(i) *Saturation regime: delegation substitutes for raw strength.* For depth budgets large enough that the orchestrator can explore most of its policy class, the per-question accuracy gap between the strong and weak base model should *shrink* relative to the no-delegation case, since both models converge towards the same evidence-bounded ceiling.

(ii) *Pre-saturation regime: delegation amplifies raw strength.* For depth budgets too small to fully exploit the policy class, the strong model extracts more value from each marginal delegation than the weak model does, so the gap should *widen* with D before saturation kicks in.

The cross-model robustness experiment (Section A.4.5) is consistent with this picture: among complete non-direct Orchestrator runs, the stronger `gpt-5.4-mini` reaches its best point at $D{=}2, T{=}2$, while `gpt-5.4-nano` and DeepSeek v4 Flash prefer the deeper $D{=}3, T{=}2$ setting. We treat this as qualitative evidence only, because the non-OpenAI delegated sweeps are still incomplete.

We test the framework's quantitative predictions in Section Q.3; the cross-model generality result above is reported in full in Section A.4.5.

H. Selective-delegation diagnostic: full results and robustness

This appendix collects the figures, sensitivity analyses, and robustness panels supporting Section A.5.

H.1. Main diagnostic details

In order to convert under-spending into a cost–quality win, the Orchestrator must spend it *selectively*: more on questions where extra compute is more likely to help, less on the rest (H4). We test this directly.

Setup. Consider the simplest version of the routing decision. After a cheap initial history H, the policy either stops (incurring conditional expected loss $L_0(H)$) or pays one extra delegation step (incurring $L_1(H)$). Define the conditional *marginal value of compute* $M(H) = L_0(H) - L_1(H)$. Selective routing in the sense of H4 means concentrating delegations on histories where $M(H)$ is large. However, the Orchestrator has *no access to the loss* at decision time: the resolution Y is

Table 8. Two-stage selective-delegation diagnostic at $D{=}2, T{=}2$. Stage 1: orchestrator LLM cost vs. direct-baseline uncertainty (positive on three of five). Stage 2: direct-baseline uncertainty vs. paired Brier improvement (positive on all five at $p \leq 0.007$)

	Stage 1: $\rho_S(u_0, \text{cost})$		Stage 2: $\rho_S(u_0, \widehat{M}^{\text{Brier}})$	
model	ρ	p (one-sided)	ρ	p (two-sided)
gpt-5.4-mini	$+0.145$	0.028	$+0.303$	< 0.001
gpt-5.4-nano	$+0.144$	0.030	$+0.220$	0.001
claude-sonnet-4.5	$+0.135$	0.042	$+0.178$	0.007
deepseek-v4-flash	-0.072	0.328	$+0.211$	0.004
gemini-2.5-flash	$+0.015$	0.826	$+0.207$	0.002

unobserved at inference, so the router cannot condition on any correctness signal, and in particular cannot route on $M(H)$ directly.

Two-stage decomposition. We decompose the mechanism into two falsifiable claims: **Stage 1.** The Orchestrator's per-question spend correlates with a computable signal that the orchestrator could plausibly use. *Does the router behave adaptively?* **Stage 2.** That signal is, in turn, correlated with the ex post marginal value of compute. *Is the router's target a valid proxy for $M(H)$?* If both hold, then the Orchestrator might be routing more compute onto questions with higher expected marginal value, even though it cannot observe $M(H)$ directly. If only Stage 2 holds, the signal is valid but the router does not use it, identifying headroom for a router that uses the signal as an explicit input.

The signal. A possible signal is the *direct-baseline uncertainty* $u_0(h) = 1 - |2p_0(h) - 1|$, where $p_0(h) = \Pr(Y = 1 \mid h)$ is the policy's posterior. u_0 is high when the cheap forecast is near 0.5 and low when the cheap policy is highly confident; it is computable from h alone, since the orchestrator runs an internal direct-style draft as part of its decision loop. We pair $D{=}0, T{=}2$ against the headline routed config $D{=}2, T{=}2$ (freeform) on the same 228 questions across the five base models, and instantiate Stage 2 as the realised paired Brier improvement $\widehat{M}^{\text{Brier}}_q = (\tilde{p}_{0,q} - y_q)^2 - (\tilde{p}_{r,q} - y_q)^2$, where $\tilde{p}_{0,q}$ is the direct baseline's ($D{=}0, T{=}2$) parsed probability forecast on question q, $\tilde{p}_{r,q}$ is the routed Orchestrator's ($D{=}2, T{=}2$) parsed forecast, and $y_q \in \{0, 1\}$ is the realised resolution. The spend variable is the orchestrator-side LLM cost, excluding web-filter cost.

Stage 1 result. Table 5 reports the Spearman correlation of u_0 with the routed Orchestrator's per-question LLM cost. Three of five models route on uncertainty at conventional significance: gpt-5.4-mini ($\rho = +0.145$, $p = 0.028$), gpt-5.4-nano ($+0.144$, $p = 0.030$), and claude-sonnet-4.5 ($+0.135$, $p = 0.042$). The remaining two models do not. Robustness across $T \in \{1, 2\}$ and $D \in \{2, 3, 5\}$.

Stage 2 result. u_0 is positively correlated with $\widehat{M}^{\text{Brier}}$ on every model, $\rho \in [0.18, 0.30]$ at $p \leq 0.007$ (Table 5, right; scatter in Appendix G.6): on questions where the cheap baseline is more uncertain, the routed forecast tends to land closer to the realised outcome than the direct one. What varies across models is therefore whether *spend* is also correlated with u_0 (Stage 1), not whether u_0 predicts paired improvement (Stage 2).

Interpretation and limitations. We emphasise that both stages are *rank correlations*, not causal claims: we do not intervene on spend, and the Orchestrator does not observe $\widehat{M}^{\text{Brier}}$ at decision time. On the three models where Stage 1 is significant, the joint pattern (spend correlates with u_0, and u_0 correlates with $\widehat{M}^{\text{Brier}}$) is consistent with the selective-delegation mechanism (H4): the Orchestrator's per-question spend is concentrated on questions where extra compute correlates ex post with larger Brier improvement. Stage 2 is positive on all five models, but Stage 1 is positive at conventional significance on only three.

Brier decomposition (lens for the informativeness check). For a forecaster that emits probabilities partitioned into bins $\{B_p\}$ with empirical bin frequency w_p, mean forecast p, and within-bin outcome rate $\bar{y}_p$ we have that:

$$\text{Brier} = \underbrace{\sum_p w_p(\bar{y}_p - p)^2}_{\text{Reliability}} - \underbrace{\sum_p w_p(\bar{y}_p - \bar{y})^2}_{\text{Resolution}} + \underbrace{\bar{y}(1-\bar{y})}_{\text{Uncertainty}}.$$

Resolution measures how informative the forecaster's confidence is about the actual outcome distribution. The Mann–Whitney test of $u_0 \mid$ wrong stochastically larger than $u_0 \mid$ right is a one-sided resolution test: informativeness can hold

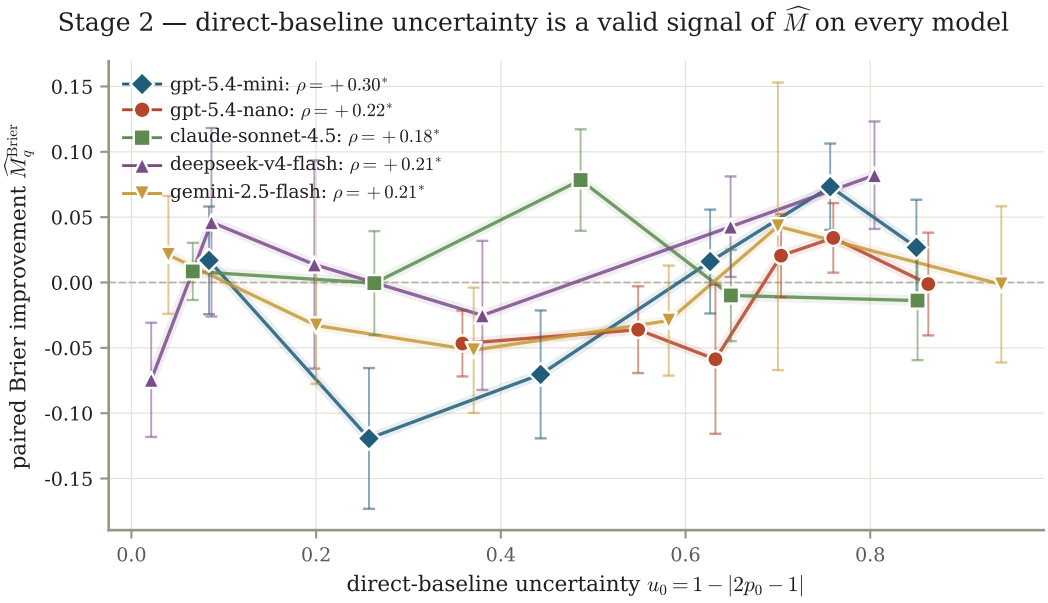


Figure 6. Stage 2 scatter (deferred from Section A.5). Paired Brier improvement $\widehat{M}_q^{\text{Brier}}$ as a function of direct-baseline uncertainty u_0, all five base models. Markers are binned means with ± 1 s.e., six bins per model. The relationship is positive on every model at $p \leq 0.007$.

whether or not the baseline is well-calibrated, and we do not test calibration here.

Robustness across routed configurations. The body restricts to the headline routed config $D{=}2, T{=}2$. Tables 9 and 10 re-run Stage 1 (cost) and Stage 2 across all routed budgets we ran: $T \in \{1, 2\}$ and $D \in \{2, 3, 5\}$, paired against matching $D{=}0$ baselines. The headline understates the effect for nano ($\rho = +0.232$, $p < 0.001$ at $D{=}3, T{=}2$) and misses a real mechanism for deepseek at the deepest budget ($\rho = +0.325$, $p < 0.05$ at $D{=}5, T{=}2$). The Stage 2 signal is essentially universal: 17 of 20 cells significant, all 20 positive in sign.

Table 9. Stage 1 (cost) across routed configurations. Spearman $\rho_S(u_0, \text{cost}_r^{\text{orch}})$ per (D, T). $^{*}\, p < 0.05$, $^{***}\, p < 0.001$.

model	$D{=}2, T{=}1$	$D{=}2, T{=}2$	$D{=}3, T{=}2$	$D{=}5, T{=}2$
gpt-5.4-mini	$+0.075$	$+0.145^{*}$	$+0.156^{*}$	$+0.047$
gpt-5.4-nano	$+0.039$	$+0.144^{*}$	$+0.232^{***}$	$+0.068$
claude-sonnet-4.5	$+0.046$	$+0.135^{*}$	$+0.129$	$+0.061$
deepseek-v4-flash	$+0.133^{*}$	-0.072	-0.052	$+0.325^{*}$
gemini-2.5-flash	-0.005	$+0.015$	$+0.013$	-0.017

Stage 1 (tokens) across routed configurations. Table 11 repeats the Stage 1 robustness sweep on completion-token count instead of cost. Cost and tokens are nearly perfectly rank-correlated within each model, and the qualitative picture is preserved: positive ρ on mini/nano/sonnet at $T{=}2$, deepseek at the extremes, gemini null throughout.

I. Dataset card

We release a dataset card following the structure recommended by the NeurIPS Datasets and Benchmarks track and by recent forecasting-benchmark work (Karger et al., 2025). The card documents:

- **Source and sourcing pipeline.** The 228 questions are sub-sampled from 12 ForecastBench question-set partitions dated `2025-10-26` through `2026-03-29`; the per-partition counts are listed in Table 12.
- **Curation constraints.** (i) Every question's `market_info_close_datetime` is strictly after the `gpt-5.4-nano`

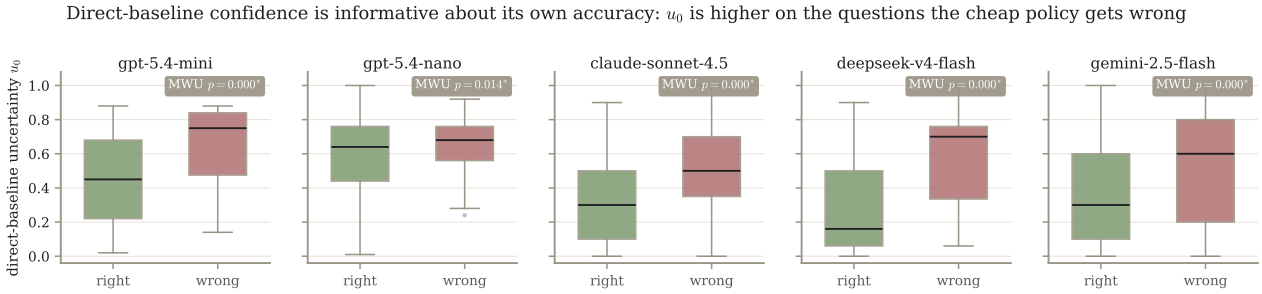


Figure 7. Direct-baseline uncertainty conditional on whether the cheap baseline gets the question right or wrong. On every model, u_0 is stochastically larger when the baseline is wrong (one-sided Mann–Whitney U, $p \leq 0.014$). This is the property that makes u_0 a usable routing signal.

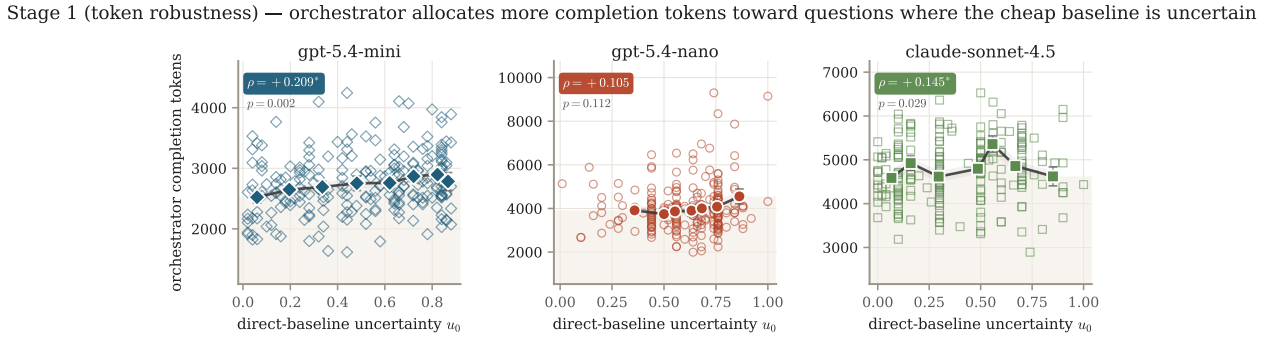


Figure 8. **Stage 1 (token robustness).** Same three panels as Figure 9-top but with completion-token count on the y-axis. Cost and tokens are nearly perfectly rank-correlated within each model; the conclusions are stable, with completion tokens giving a slightly stronger signal for mini ($\rho = +0.209$, $p = 0.001$).

knowledge cutoff (∼September 2025), with resolution dates February 2026–January 2027; (ii) the resolved label distribution is exactly 114 True / 114 False. The original ForecastBench partitions are False-heavy, so during curation we balance the set by including polarity-flipped variants of some questions: when a source question resolves False, its negated proposition is included as a True-resolved question where the wording remains natural and the resolution is unambiguous; (iii) topical mix is preserved from the underlying ForecastBench partitions (Politics/Geopolitics 49, Technology 32, Sports 31, Economics/Finance 21, Entertainment 5, Science/Climate 3, Other/Mixed 87).

- **Reproducibility metadata.** We retain the source ForecastBench identifiers, partition labels, resolution dates, topics, ground-truth labels, and polarity-flip indicators in the released question file, and credit ForecastBench as the source benchmark.

J. Side experiment: pre-cutoff questions (contamination-control validation)

The headline study (Section A.4) deliberately uses post-knowledge-cutoff questions and a contamination-filtering web layer because pre-cutoff data is at risk of leakage through training-time exposure or through unfiltered retrieval of post-resolution content (Paleka et al., 2025). The severity of this risk is confirmed by Li et al. (2026) (arXiv:2601.13717), who show across 477 competition-level questions and 9 models that prompting an LLM to "ignore" knowledge acquired before its cutoff fails systematically: a 52% performance gap persists between simulated ignorance and true ignorance, and chain-of-thought reasoning *cannot* suppress prior knowledge even when reasoning traces contain no explicit post-cutoff references. This finding directly motivates the design of our pre-cutoff side experiment: we run it not to *use* pre-cutoff data as a fair test, but to diagnose whether the architecture-level allocation pattern persists under a different contamination regime. We therefore focus on qualitative system behaviour, namely which architectures spend compute, call tools, and sit on the cost–quality frontier, rather than interpreting absolute accuracy as forecasting skill. A reviewer could reasonably worry that the *architectural*

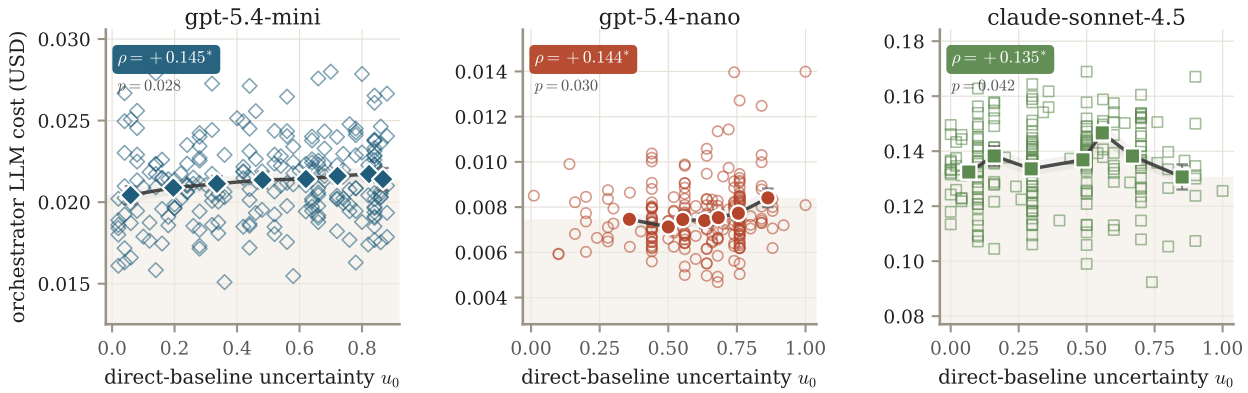


Figure 9. **Stage 1.** Orchestrator LLM cost vs. direct-baseline uncertainty u_0 for the three models with positive Spearman correlation ρ . Markers are binned means with ± 1 s.e.

Table 10. Stage 2 across routed configurations. Spearman $\rho_S(u_0, \widehat{M}^{\text{Brier}})$ per (D, T). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. 17 of 20 cells significant.

model	$D{=}2, T{=}1$	$D{=}2, T{=}2$	$D{=}3, T{=}2$	$D{=}5, T{=}2$
gpt-5.4-mini	$+0.180^{**}$	$+0.303^{***}$	$+0.219^{**}$	$+0.094$
gpt-5.4-nano	$+0.078$	$+0.220^{**}$	$+0.219^{**}$	$+0.123$
claude-sonnet-4.5	$+0.143^{*}$	$+0.178^{**}$	$+0.335^{***}$	$+0.253^{***}$
deepseek-v4-flash	$+0.231^{***}$	$+0.211^{**}$	$+0.169^{*}$	$+0.617^{***}$
gemini-2.5-flash	$+0.220^{*}$	$+0.207^{**}$	$+0.175^{**}$	$+0.156$

ranking we report, where orchestrator Pareto-dominates breadth and depth, is itself an artefact of how the post-cutoff + filtered setting interacts with each architecture. This side experiment is designed to falsify that worry.

J.1. Setup

We re-run the *exact* same code base and configuration sweep used in the main study, on a separate set of 72 ForecastBench questions dated `2024-12-08` whose resolution dates are *before* the model knowledge cutoff. The base model (`gpt-5.4-nano`), tools, web filter, midpoint cutoff convention, and analysis pipeline are identical; only the question set differs. This isolates the effect of the contamination-control regime on the architectural ranking and on each architecture's scaling behaviour.

J.2. What we test

1. **Architectural ranking preservation**: does the orchestrator still Pareto-dominate ensemble and predictor–critic on cost, tokens, and LLM calls?

2. **Diminishing-returns shape preservation**: does the marginal-accuracy-per-extra-dollar curve along each architecture's efficient frontier still drop sharply after one or two points?

3. **Retrieval-use persistence**: do agents still spend test-time compute on tool use and retrieval even when questions resolve before the model cutoff? The web-filter cost share is a proxy for this behaviour because it is incurred when agents retrieve pages that must be screened.

4. **Calibration profile**: do all architectures still exhibit non-trivial ECE, and is the headline post-hoc recovery procedure (Platt or temperature scaling) still effective in the contaminated regime?

J.3. Results

The four checks above are preserved on the side benchmark (Figures 10 to 12), and the architectural ranking is in fact *stronger* on the pre-cutoff set than on the headline set:

Table 11. Stage 1 (tokens) across routed configurations. Spearman $\rho_S(u_0, \text{completion tokens}_r)$ per (D, T). $^{*}\, p < 0.05$, $^{**}\, p < 0.01$.

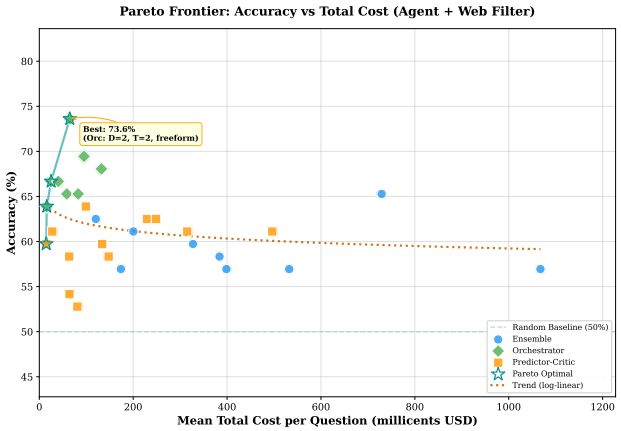
model	$D{=}2, T{=}1$	$D{=}2, T{=}2$	$D{=}3, T{=}2$	$D{=}5, T{=}2$
gpt-5.4-mini	$+0.087$	$+0.209^{**}$	$+0.186^{**}$	$+0.094$
gpt-5.4-nano	$+0.028$	$+0.105$	$+0.205^{**}$	$+0.076$
claude-sonnet-4.5	$+0.037$	$+0.145^{*}$	$+0.126$	$+0.071$
deepseek-v4-flash	$+0.163^{*}$	-0.028	-0.000	$+0.367^{*}$
gemini-2.5-flash	-0.010	$+0.029$	$+0.008$	-0.001

Table 12. Per-partition counts of the curated 228-question benchmark. Partitions are the underlying ForecastBench question-set release dates; the curation jointly imposes balanced T/F at the dataset level and post-cutoff resolution at the question level.

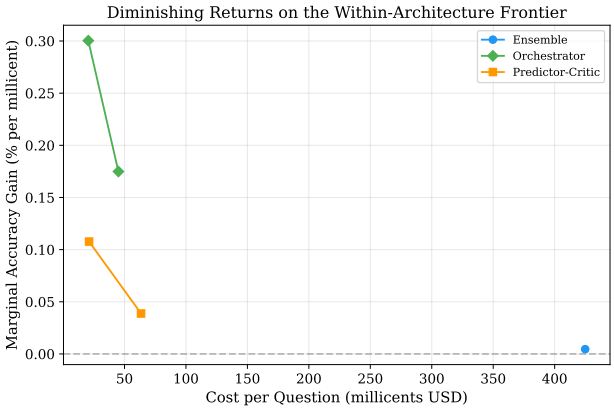
Partition (date)	N	#True	#False
2025-10-26	34	20	14
2025-11-09	28	10	18
2025-11-23	18	7	11
2025-12-07	23	12	11
2025-12-21	21	6	15
2026-01-04	10	5	5
2026-01-18	8	3	5
2026-02-01	9	4	5
2026-02-15	10	5	5
2026-03-01	24	16	8
2026-03-15	20	13	7
2026-03-29	23	13	10
Total	**228**	**114**	**114**

- *Pareto ranking.* The orchestrator owns the entire cost–accuracy frontier on the side experiment. Best configurations: Orchestrator $D{=}2, T{=}2$ (freeform) at 73.6% **/ $0.065/q**; Ensemble $K{=}4, R{=}2$ at 65.3% / \$0.729/q; Predictor–Critic $T{=}1, E{=}2$ at 63.9% / \$0.099/q. Same ordering, larger raw gap ($\sim$ 8–10 percentage points to either static policy).
- *Diminishing returns.* The marginal accuracy gain per extra dollar again drops by more than an order of magnitude after the first one or two frontier points, matching the headline shape.
- *Retrieval-use persistence.* The contamination-filter cost share at the best configuration of each architecture is 89.2% (Orchestrator), 88.9% (Predictor–Critic), and 90.1% (Ensemble); across the full 29-config side sweep the share spans 85.7–95.2% (median 90.7%). This is not important because the filter share is intrinsically interesting; rather, it shows that agents still make tool calls and engage the retrieval pipeline even in a regime where some outcome-relevant information may already be present in model weights.
- *Calibration is non-trivial but qualitatively different in direction.* All best-per-architecture configurations exhibit non-trivial raw ECE: 0.229 (Ens), 0.174 (Orch), and 0.170 (PC), substantially higher than the headline $[0.09, 0.16]$ band, consistent with residual training-time leakage that biases the model toward confident-but-not-always-right answers on pre-cutoff questions. Post-hoc recovery is also more variable than on the headline: Platt scaling reduces ECE by 60.9% on Ensemble and 16.2% on Orchestrator, but *worsens* ECE by 5.4% on Predictor–Critic; temperature scaling is the more reliable recovery method on the side benchmark (53.0% / 12.7% / 25.3% reduction across the three architectures). We treat this as expected: the calibration-recovery method best-matched to the paper's headline regime (two-parameter Platt on under-confident forecasts) is not necessarily best-matched to the contaminated regime where the bias structure of the raw forecasts is different. The calibration profile is therefore qualitatively non-trivial in both regimes but should be analysed regime-by-regime, not assumed to transfer.
- *Statistical significance.* As on the headline study, bootstrap CIs straddle zero at $N{=}72$ (Orch vs. Ens: $\Delta\text{Acc} = +0.083, p = 0.11$; Orch vs. PC: $\Delta\text{Acc} = +0.097, p = 0.07$). The robust claim remains cost–quality dominance.

The slightly larger architectural gap on the side benchmark is consistent with the residual training-time leakage diagnosed by Li et al. (2026); Lahib et al. (2026): the orchestrator's ability to redirect retrieval evidence selectively may be even more useful when the underlying retrieval surface and parametric memory are partially contaminated. The validity claim the side

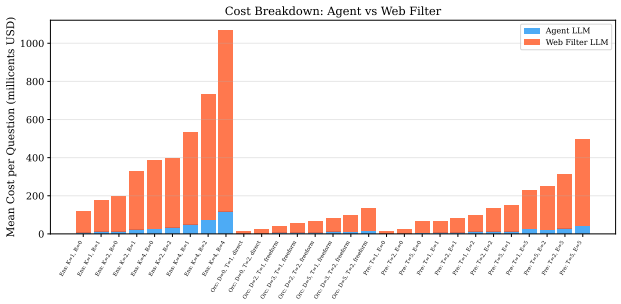


(a) Pareto: accuracy vs. total cost (72-Q pre-cutoff)

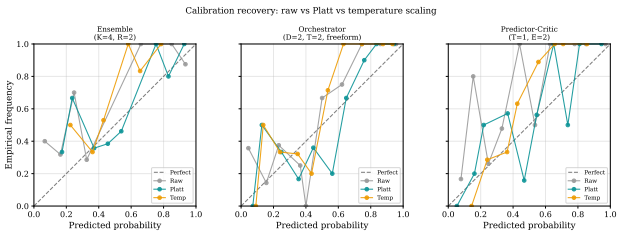


(b) Diminishing returns along the efficient frontier (72-Q)

Figure 10. Side experiment: architectural ranking and the diminishing-returns shape are preserved on the 72-question pre-cutoff set; the headline finding is not specific to the post-cutoff regime.



(a) Cost decomposition (72-Q pre-cutoff)



(b) Calibration recovery: raw vs. Platt vs. temperature (72-Q)

Figure 11. Agents continue to engage the retrieval pipeline on pre-cutoff questions: the web-filter cost share at the best configuration of each architecture remains high (89.2% Orch, 88.9% PC, 90.1% Ens). The side-experiment raw ECE values are higher than on the headline benchmark (0.17–0.23 vs 0.09–0.16), and post-hoc calibration is more variable: Platt scaling helps Ensemble and Orchestrator but hurts Predictor–Critic, while temperature scaling is more reliable across all three architectures.

experiment therefore supports is the *qualitative* preservation of the architecture-level allocation and retrieval-use patterns reported in the headline study, not a quantitative replacement of the headline numbers.

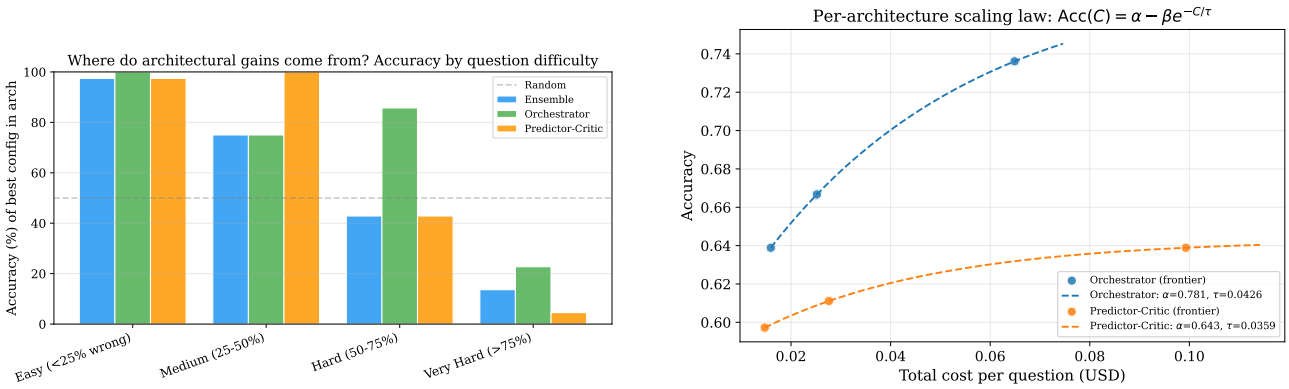
J.4. Where the side experiment differs

Two differences from the main study are expected and observed:

1. **Absolute accuracy is higher on the pre-cutoff set** for every architecture. We do *not* treat this as a positive result for the model: it is consistent with residual leakage through training-time exposure that the web filter cannot block, and is one of the reasons the headline study uses post-cutoff questions. Pre-cutoff exposure does not imply perfect accuracy: outcome-relevant facts may be absent from training data, not reliably retrieved from model weights, obscured by question wording or polarity flips, or blocked from retrieval by the web filter. We therefore interpret higher pre-cutoff accuracy as evidence of possible exposure, not as evidence that the task is trivial.
2. **Statistical power is lower** (N=72 vs. N=228); the side experiment is intended to validate *patterns*, not to provide a competitive accuracy number.

J.5. Take-away

The *architectural ranking*, the *diminishing-returns shape*, and the agents' *retrieval-use behaviour* are all preserved on a separate 72-question pre-cutoff benchmark using the same model and pipeline.



(a) Difficulty stratification (72-Q) *(b)* Saturating scaling-law fits (72-Q)

Figure 12. Difficulty stratification and the saturating scaling-law fit on the 72-Q side experiment.

K. Full per-configuration tables

This appendix reports the full configuration grids underlying Table 1. We include accuracy, Brier score, total pipeline cost, and agent-side LLM calls; the released `metrics_by_config.csv` files include additional token, agent-only cost, web-filter cost, and utilisation columns.

Table 13. Full corrected `gpt-5.4-mini` configuration results.

Architecture	Config	N	Acc. (%)	Brier	Cost ($/q)
Ensemble	K=1, R=1	228	73.7	0.193	0.6922
Ensemble	K=1, R=2	228	78.1	0.180	0.8970
Ensemble	K=1, R=4	228	75.4	0.178	1.2412
Ensemble	K=4, R=1	228	76.8	0.181	1.9585
Ensemble	K=4, R=2	228	78.1	0.178	2.6631
Ensemble	K=4, R=4	226	77.9	0.172	3.6091
Orchestrator	D=0, T=1 direct	228	73.7	0.199	0.0469
Orchestrator	D=0, T=2 direct	228	76.3	0.178	0.0751
Orchestrator	D=2, T=1 freeform	228	77.6	0.183	0.1093
Orchestrator	D=3, T=1 freeform	228	74.6	0.207	0.1432
Orchestrator	D=2, T=2 freeform	228	80.7	0.187	0.1752
Orchestrator	D=5, T=1 freeform	228	75.4	0.194	0.2016
Orchestrator	D=3, T=2 freeform	228	75.4	0.211	0.2606
Orchestrator	D=5, T=2 freeform	228	78.1	0.190	0.3357
Predictor–Critic	T=2, E=0	228	64.5	0.224	0.0971
Predictor–Critic	T=5, E=0	228	68.9	0.201	0.2106
Predictor–Critic	T=2, E=1	228	71.1	0.199	0.2826
Predictor–Critic	T=2, E=2	228	72.4	0.185	0.4736
Predictor–Critic	T=5, E=1	228	72.8	0.195	0.5323
Predictor–Critic	T=5, E=2	228	75.4	0.169	0.7708
Predictor–Critic	T=2, E=5	228	76.3	0.169	0.9398
Predictor–Critic	T=5, E=5	228	76.8	0.168	1.6649

L. Additional Pareto views

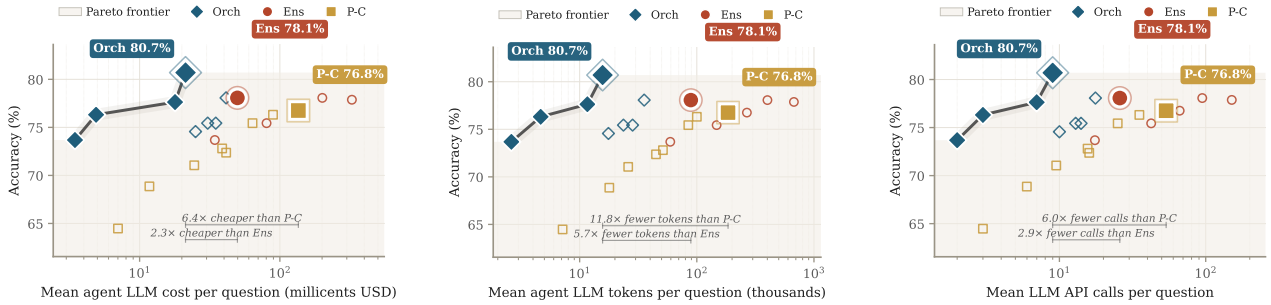
The main results use total pipeline cost because the web-filtering layer is part of the contamination-controlled backtesting pipeline (Figure 2). To verify that the architecture ranking is not an artefact of web-filter pricing, Figure 13 repeats the Pareto comparison using three agent-side or pricing-independent proxies: agent-only LLM cost, agent-token consumption, and LLM API calls. In all three views, the observed Pareto frontier is again composed of Orchestrator configurations only. Thus H1 is supported both in the full backtesting-cost accounting and in the agent-only compute accounting.

M. Full architecture-grid heatmaps

Figure 14 gives the full grid behind the saturation and non-monotonicity claim in Section A.4.2. The heatmaps show that larger static schedules do not reliably improve accuracy. In the Ensemble grid, increasing from the cheapest K=1, R=1

Table 14. Full `gpt-5.4-nano` configuration results.

Architecture	Config	N	Acc. (%)	Brier	Cost ($/q)
Ensemble	K=1, R=1	228	67.5	0.211	0.1566
Ensemble	K=1, R=2	228	68.4	0.193	0.2027
Ensemble	K=1, R=4	228	70.2	0.196	0.2942
Ensemble	K=4, R=1	228	70.2	0.195	0.4620
Ensemble	K=4, R=2	228	72.4	0.183	0.6086
Ensemble	K=4, R=4	228	72.8	0.179	0.8992
Orchestrator	D=0, T=1 direct	228	64.0	0.222	0.0111
Orchestrator	D=0, T=2 direct	228	68.9	0.209	0.0161
Orchestrator	D=2, T=1 freeform	228	64.9	0.226	0.0314
Orchestrator	D=3, T=1 freeform	228	65.4	0.224	0.0439
Orchestrator	D=2, T=2 freeform	228	68.4	0.221	0.0494
Orchestrator	D=5, T=1 freeform	228	68.9	0.218	0.0628
Orchestrator	D=3, T=2 freeform	228	71.1	0.209	0.0699
Orchestrator	D=5, T=2 freeform	228	70.2	0.211	0.1075
Predictor–Critic	T=1, E=0	228	61.0	0.234	0.0122
Predictor–Critic	T=2, E=0	228	63.6	0.236	0.0246
Predictor–Critic	T=1, E=1	228	65.4	0.230	0.0471
Predictor–Critic	T=5, E=0	228	63.2	0.225	0.0475
Predictor–Critic	T=2, E=1	228	63.6	0.218	0.0695
Predictor–Critic	T=1, E=2	228	63.6	0.230	0.0817
Predictor–Critic	T=2, E=2	228	64.0	0.212	0.1172
Predictor–Critic	T=5, E=1	228	68.9	0.207	0.1248
Predictor–Critic	T=1, E=5	228	68.0	0.210	0.1803
Predictor–Critic	T=5, E=2	228	71.5	0.194	0.1894
Predictor–Critic	T=2, E=5	228	70.2	0.202	0.2577
Predictor–Critic	T=5, E=5	228	68.9	0.209	0.4027



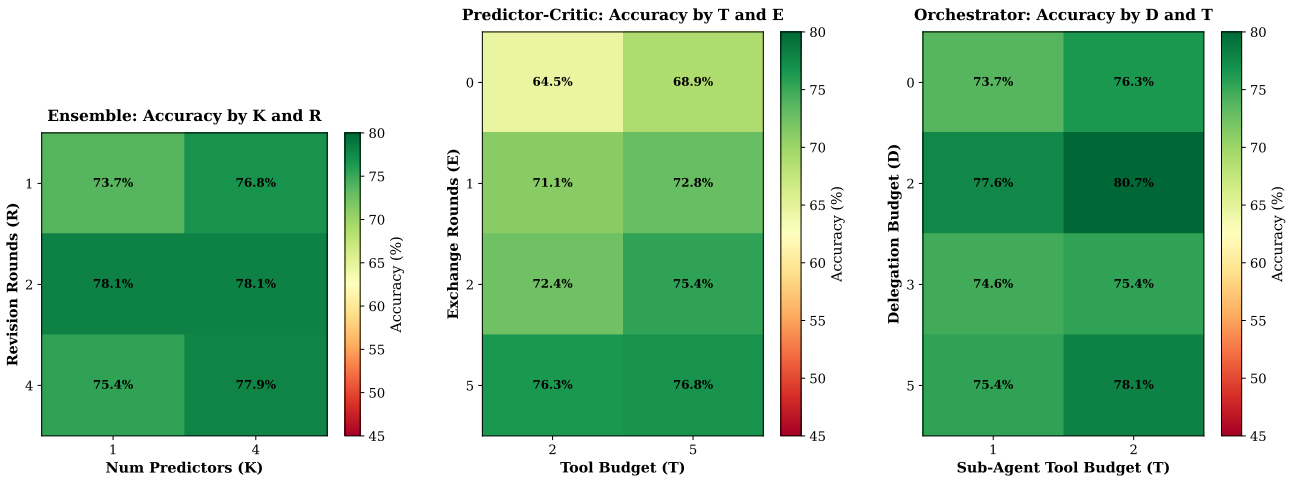
(a) Agent-only cost / question *(b)* Agent tokens / question *(c)* LLM API calls / question

Figure 13. Pareto views under alternative compute proxies. These reproduce the cost–accuracy ordering from Figure 2, showing that the result is not an artefact of web-filter pricing, dollar pricing, or token accounting.

setting to larger K, R can help, but the high-cost K=4, R=4 cell does not beat the cheaper K=1, R=2 or K=4, R=2 cells. In Predictor–Critic, increasing exchange depth and tool budget helps up to a point, after which additional compute yields small or non-monotone gains. In the Orchestrator grid, D=2, T=2 is the best cell; larger delegation budgets do not improve accuracy. This supports H2 in the empirical sense used in the main text: accuracy saturates or becomes non-monotone as budget grows, rather than increasing smoothly with raw test-time compute.

N. Architecture-router oracle analysis

The best configuration of each architecture solves a partially different subset of questions. On the `mini` sweep, the best Ensemble (K=1, R=2) answers 178/228 questions correctly, the best Orchestrator (D=2, T=2) answers 184/228, and the best Predictor–Critic (T=5, E=5) answers 175/228. An oracle that selects the correct architecture per question among these three best configurations would answer 212/228 questions correctly (93.0%). Among the 228 questions, all three architectures are correct on 142, all three are wrong on 16, exactly one is correct on 29, and exactly two are correct on 41. The unique-correct contributions are 8 for Ensemble, 16 for Orchestrator, and 5 for Predictor–Critic.



(a) Ensemble: $K \times R$ *(b)* Pred.–Critic: $T \times E$ *(c)* Orchestrator: $D \times T$

Figure 14. Accuracy heatmaps over each architecture's parameter grid. These provide full-grid transparency for the diminishing-returns result in Figure 3.

Table 15. Cross-architecture complementarity among the best `mini` configurations. "Oracle" is an upper bound: it selects the correct architecture after observing the label.

Selector	Correct	Acc. (%)	Notes
Best single config (Orchestrator)	184/228	80.7	D=2, T=2
Oracle over best three architectures	212/228	93.0	Uses realised labels
Oracle over all Ensemble configs	208/228	91.2	All K, R cells
Oracle over all Orchestrator configs	214/228	93.9	All D, T cells
Oracle over all Predictor–Critic configs	206/228	90.4	All T, E cells
Oracle over all reported `mini` configs	223/228	97.8	All reported cells

O. Calibration and agreement details

Per-architecture calibration recovery. Table 16 reports the full Platt and temperature-scaling output for the best configuration of each architecture, both in-sample (fit on all 228 questions, evaluated on the same data) and held-out (stratified 5-fold cross-validation, evaluated on the concatenated out-of-fold predictions). The held-out numbers should be trusted for a deployment claim; we report the in-sample numbers as well so that readers can verify that the calibrator is not strongly over-fitting the small 228-question pool. The in-sample and held-out ECE values are close across the three best configurations; the largest gap is 2.5 pp for Ensemble under Platt scaling, reflecting the greater variability of its predictions.

Table 16. Calibration recovery on the best configuration of each architecture: Platt (two parameters) and temperature scaling (one parameter) against the raw posterior, with bin-wise expected (ECE) calibration error. Each calibrator is reported both *in-sample* (fit and evaluated on all 228 questions) and *held-out* (5-fold stratified cross-validation; ECE computed on the concatenated out-of-fold predictions). The cross-validated Platt parameters are reported as $(\bar{a} \pm \sigma_a,\ \bar{b} \pm \sigma_b)$ across folds.

	$\text{ECE}_{\text{Platt}}$		$\text{ECE}_{\text{temp.}}$		Δ Platt		
Architecture	in-s.	held-out	in-s.	held-out	in-s.	held-out	Platt (a, b) (CV)
Ensemble K=1, R=2	0.107	0.082	0.100	0.107	−25%	−42%	$(0.80\pm0.06,\ 0.79\pm0.10)$
Predictor–Critic T=5, E=5	0.087	0.078	0.108	0.107	−33%	−39%	$(1.05\pm0.08,\ 0.87\pm0.10)$
Hierarchical Orch. D=2, T=2	0.083	0.087	0.107	0.098	−37%	−34%	$(0.66\pm0.05,\ 0.27\pm0.07)$
Reference: ECE_{raw} = 0.142 / 0.129 / 0.131 for Ensemble / PC / Orchestrator (same row order).							

One plausible explanation is that the static architectures induce more conservative probability reports: Ensemble aggregation averages across independent predictors, and Predictor–Critic critique can temper an initial forecast before the final answer. The Orchestrator, by contrast, often synthesizes a single routed evidence trajectory, which may produce sharper but less

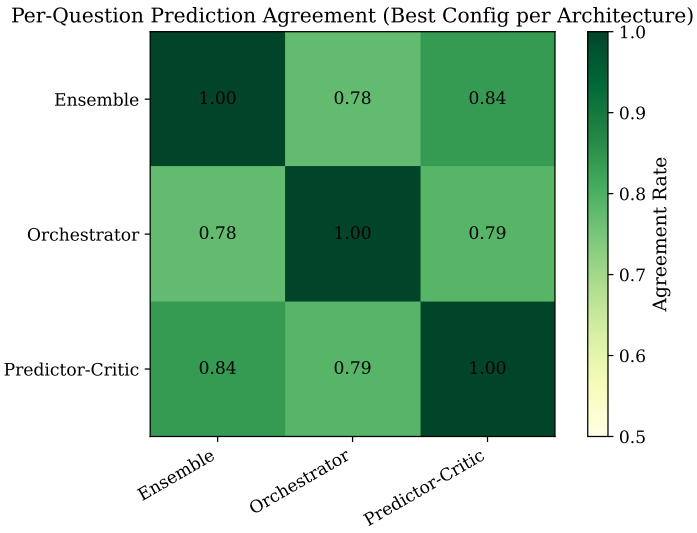
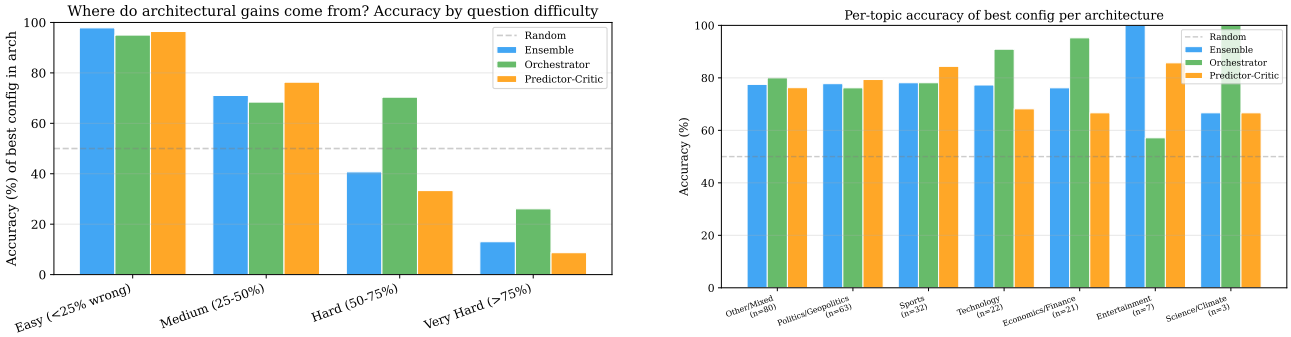


Figure 15. Pairwise binary-prediction agreement between the best-per-architecture configurations on the 228-question benchmark. Off-diagonal entries lie in $[77.2\%, 79.4\%]$, meaning the three architectures disagree on roughly 20% of questions. This is the headroom targeted by the architecture-routing direction discussed in Section A.6.

calibrated raw probabilities. We treat this as a calibration hypothesis rather than a proven mechanism; the post-hoc results show that much of the calibration gap is recoverable with simple scaling.

P. Difficulty and topic stratification



(a) Accuracy by difficulty band

(b) Accuracy by topic

Figure 16. Where architectures differ across question strata. Difficulty bands are defined by across-architecture consensus error rate; topic labels come from the benchmark metadata.

Q. Theoretical-framework details

Q.1. Formal statement and proof of Proposition G.1

Setup. Fix a question–label distribution $(Q, Y) \sim \mathcal{D}$, a finite action set $\mathcal{A}$, a per-action unit cost $c : \mathcal{A} \to \mathbb{R}_{\geq 0}$, and a budget $C \geq 0$. A *compute-bounded policy* is a (possibly randomised) decision rule π that selects $a_t \in \mathcal{A} \cup \{\text{TERMINATE}\}$ from (Q, h_t), with $h_t = (a_1, o_1, \ldots, a_{t-1}, o_{t-1})$ the realised history of actions and observations, and that respects $\sum_t c(a_t) \leq C$ almost surely. On termination π emits $(\hat{y}, \hat{p}) \in \{0, 1\} \times [0, 1]$. Let $\Pi_{\text{static}} = \{\pi : a_t \perp h_t \mid Q, t\}$ and Π_{adapt} be all such policies. Let $L(\pi, C) = \Pr_{(Q,Y)\sim\mathcal{D}}[\hat{y} \neq Y]$.

Proposition (Restated)**.** *For all $C \geq 0$, $\min_{\pi \in \Pi_{adapt}} L(\pi, C) \leq \min_{\pi \in \Pi_{static}} L(\pi, C)$. The inequality is strict whenever there exists a measurable event $E \subset \Omega$ with $\Pr[E] > 0$ on which the optimal Bayes action under $h_t \in E$ differs from the*

optimal Bayes action under $h_t \notin E$.

Proof. Containment is immediate: $\Pi_{\text{static}} \subset \Pi_{\text{adapt}}$, so the minimum over the larger class is at least as small.

For strictness, let π_s^* achieve the static minimum and let $a_t^*(h_t)$ denote the Bayes-optimal action at step t given h_t. On E the static policy must, by definition, take an action independent of h_t, hence with positive probability strictly suboptimal under the 0–1 loss. Construct π_a as: identical to π_s^* off E, and equal to $a_t^*(h_t)$ on E. By construction π_a uses no more budget than π_s^* in expectation ($c(a_t^*) \le c(a_t^{\pi_s^*})$ by optimality of a_t^*); and $\Pr[\hat{y}_{\pi_a} \neq Y \mid E] < \Pr[\hat{y}_{\pi_s^*} \neq Y \mid E]$ by Bayes-optimality. Marginalising over E vs. $\bar{E}$ yields $L(\pi_a, C) < L(\pi_s^*, C)$. □

Remarks. (i) The proof is a one-step policy-improvement argument and does not require any specific form for the action distribution. (ii) The strict-inequality condition is mild: it asks only that the realised history sometimes carries information about the optimal next action, which is precisely what the orchestrator's freeform mode is designed to exploit. (iii) The proposition does not give a rate; the empirical rate is what the τ_{arch} in (4) measures.

Q.2. Bayes-risk derivation of the evidence ceiling

We use the standard binary Bayes-risk bound rather than Fano's inequality. For binary Y, Fano gives $H(P_e) + P_e \log(|\mathcal{Y}| - 1) \ge H(Y \mid \hat{y})$, which with $|\mathcal{Y}| = 2$ collapses to $H(P_e) \ge H(Y \mid \hat{y})$. The further bound $H(P_e) \le 1$ yields $P_e \ge H(Y \mid \hat{y}) - 1$, but this is non-positive whenever $H(Y \mid R(C)) \le 1$, which is always the case for a binary outcome, so the bound is vacuous in our setting. We therefore use the (sharper, valid, and standard) decision-theoretic bound instead.

Let $\eta(R) = \Pr(Y = 1 \mid R)$. The minimum probability of error of *any* decision rule that bases $\hat{y}$ on the evidence set R is the Bayes risk $L^*(R) = \mathbb{E}_R\big[\min(\eta(R),\, 1 - \eta(R))\big]$, and any policy π acting on R satisfies $\Pr(\hat{y} \neq Y) \ge L^*(R)$. With $R = R(C)$, we obtain $L_{\text{ev}}(C) \ge L^*(R(C))$, the bound used in Section G.4. The implication that matters for the paper's argument is qualitative: additional reasoning compute on a fixed evidence set cannot reduce $L^*(R(C))$; only compute that changes the evidence (more retrieval, broader queries) can lower the floor.

Q.3. Scaling-law fit procedure

We fit (4) to each architecture's efficient frontier using non-linear least squares with the Levenberg–Marquardt algorithm. The fitting target is mean accuracy across the 228-question benchmark, the regressor is total cost per question (USD), and we restrict to non-dominated frontier points to avoid contamination by clearly-suboptimal configurations. Confidence intervals are computed by a non-parametric bootstrap (1 000 resamples of frontier points; 2.5–97.5 percentile interval). The fitting script produces Table 17 and Figure 17.

Why the parametric form. The saturating exponential is the unique form implied by a first-order relaxation $d\,\text{Acc}/dC = (\alpha - \text{Acc}(C))/\tau$, i.e. each extra dollar of compute closes a fixed fraction of the remaining headroom to the evidence ceiling. The bounds $(\alpha, \beta, \tau) \in [0,1] \times [0,1] \times [10^{-6}, 10^{6}]$ make the fitted curve automatically monotone in cost and concave (diminishing returns), so neither has to be enforced post hoc.

Why the frontier filter. Restricting to non-dominated frontier points aligns the fitting data with these structural assumptions: a dominated configuration violates monotonicity by definition (lower accuracy at higher cost) and would otherwise force the optimiser to compromise between mutually inconsistent constraints. Empirically, 3/8 orchestrator configurations and 5/12 predictor–critic configurations sit on the Pareto frontier; all 5/5 ensemble configurations do (the breadth and revision axes both monotonically increase cost and accuracy in our sweep).

Robustness checks. We report three robustness analyses in Appendix R: (i) *all-configurations* least-squares fit (no frontier filter), which inflates τ_{Orch} by $\sim 50\times$ as expected when monotonicity-violating points are admitted; (ii) a *per-question Bernoulli MLE* that replaces the implicit Gaussian-on-aggregated-accuracies likelihood with the correct $\text{correct}_i \sim \text{Bernoulli}(\alpha - \beta e^{-c_i/\tau})$ observation model and recovers Hessian-based Wald 95 % CIs. This agrees with the frontier fit for ensemble (where the per-question signal is large) but reveals that predictor–critic's τ is weakly identified at the per-question level (saturation is supported only by the frontier-aggregated reading); (iii) replacing the exponential saturation by a power-law $\alpha - \beta C^{-\gamma}$ and replacing total cost by total tokens or total LLM calls. The qualitative ordering $\tau_{\text{Orch}} \ll \tau_{\text{PC}} < \tau_{\text{Ens}}$ is preserved across all three sensitivities.

Table 17. Fitted parameters of $\mathrm{Acc}(C) = \alpha - \beta e^{-C/\tau}$ per architecture on the `gpt-5.4-mini` efficient frontier, with non-parametric bootstrap 95% intervals. τ is in USD per question.

Architecture	α (ceiling)	β (headroom)	τ ($/q)
Predictor–Critic	0.768 [0.746, 0.857]	0.159 [0.120, 0.318]	0.330 [0.187, 1.36]
Ensemble	0.774 [0.768, 1.000]	1.000 [–, –]	0.199 [0.185, 1874]
Hierarchical Orchestrator	0.848 [0.784, 1.000]	0.156 [0.105, 0.279]	**0.132** [0.035, 0.492]

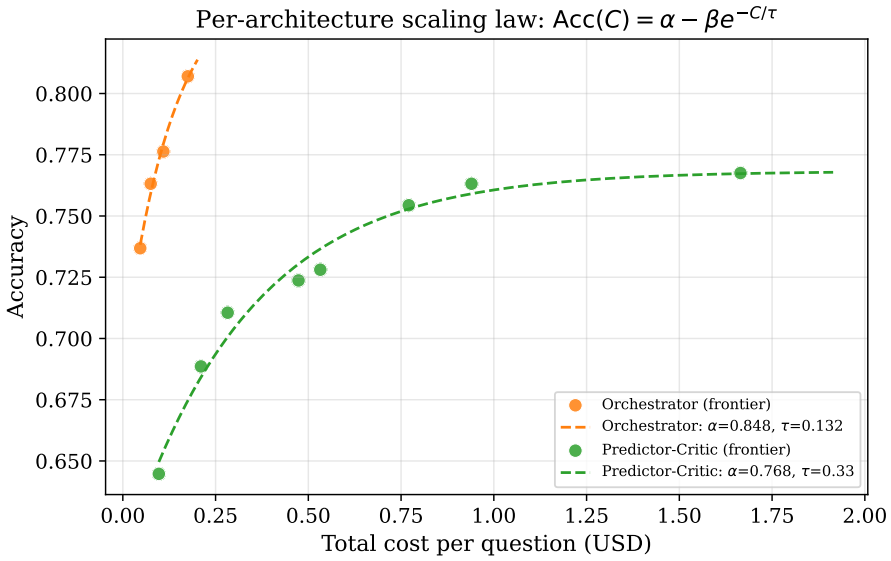


Figure 17. Per-architecture efficient frontiers with the fitted saturating model overlaid. The fits are descriptive: each frontier contains only a small number of non-dominated points and confidence intervals are correspondingly wide.

R. Scaling-law robustness

We re-fit the saturating scaling law of Section Q.3 under three perturbations of the fitting strategy and report the resulting (α, β, τ) in Table 18. The headline ordering $\tau_{\text{Orch}} \ll \tau_{\text{PC}} < \tau_{\text{Ens}}$ is preserved across all three checks; only the absolute scale of τ_{Orch} shifts when we admit dominated configurations into the fit.

Table 18. Three fits of $\mathrm{Acc}(C) = \alpha - \beta e^{-C/\tau}$ per architecture on the 228-question balanced post-cutoff set (`gpt-5.4-mini`). *frontier-LSQ* is the headline procedure used in Table 17. *all-LSQ* drops the Pareto filter; *Bernoulli MLE* replaces the Gaussian-on-aggregated likelihood with an explicit per-question Bernoulli observation model.

Architecture	Mode	N_{obs}	α	β	τ ($/q)
Ensemble	all-LSQ	6	0.774	1.000	0.199
Ensemble	Bernoulli MLE	1366	0.774	1.000	0.198
Orchestrator	frontier-LSQ	4	0.848	0.156	0.132
Orchestrator	all-LSQ	8	0.770	0.641	0.016
Orchestrator	Bernoulli MLE	1824	0.770	0.644	0.016
Predictor-Critic	frontier-LSQ	8	0.768	0.159	0.330
Predictor-Critic	all-LSQ	8	0.768	0.159	0.330
Predictor-Critic	Bernoulli MLE	1824	0.769	0.159	0.336

Interpretation. For ensemble, every configuration is on the Pareto frontier and the per-question signal is modest, so the bounded exponential fit is best read as a compact descriptive curve rather than a parameter estimate. For orchestrator, including dominated configurations changes the fitted time constant sharply because those points lie at higher cost without higher accuracy. Predictor–Critic is more stable in the `mini` sweep because all complete configurations lie on the frontier used by the fit. We therefore report the frontier-LSQ values as headline descriptive summaries, accompanied by the explicit

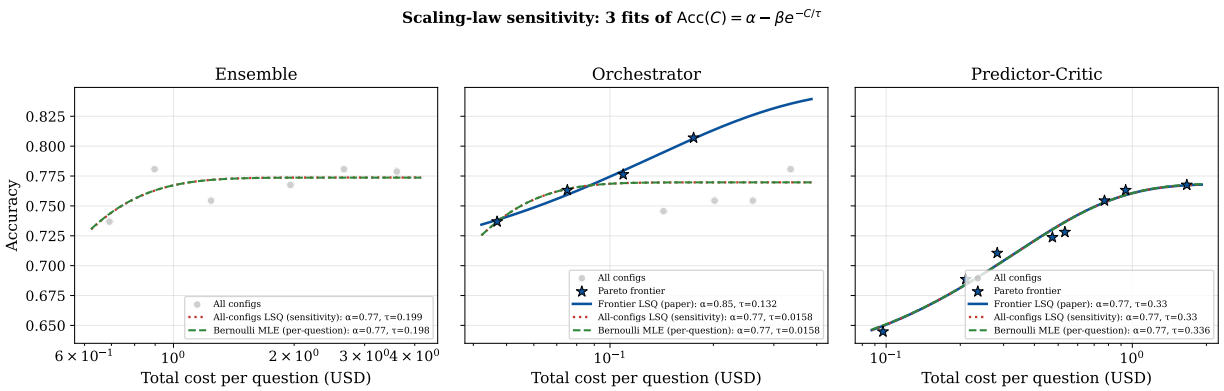


Figure 18. Sensitivity of the scaling-law fit. For each architecture we overlay the three fitting modes from Table 18. Light-grey markers are all observed configurations; navy stars mark the Pareto-optimal subset used by the headline fit. The headline (blue solid), all-configurations (red dotted), and Bernoulli-MLE (green dashed) curves agree closely for ensemble, confirming the architecture's frontier is essentially unaffected by fitting choice. They diverge for orchestrator: the all-configs and Bernoulli fits are pulled toward the dominated configurations and overestimate τ by roughly two orders of magnitude. Predictor-critic's Bernoulli fit collapses to a flat line because the per-question Bernoulli noise ($\approx$0.5) dominates the inter-config accuracy spread ($\approx$11pp); τ is therefore weakly identified at the per-question level even when it is well-determined on the frontier-aggregated curve.

caveats above.

S. Experimental setup details

Models. All agents (predictor, critic, ensemble member, aggregator, sub-agents, and orchestrator) use either `gpt-5.4-nano` or `gpt-5.4-mini` via the OpenAI Responses API with reasoning effort `medium`; the web-filter LLM is held fixed. The main body reports the `mini` sweep as the primary complete architecture comparison, and uses `nano` as a full-sweep replication. We additionally report an Orchestrator-only cross-model check on complete delegation-enabled runs from Claude Sonnet 4.5, DeepSeek v4 Flash, Gemini 2.5 Flash, `gpt-5.4-mini`, and `gpt-5.4-nano` (Section A.4.5).

Configuration sweep. For the main `mini` sweep, Predictor–Critic uses $T \in \{2, 5\}$ and $E \in \{0, 1, 2, 5\}$ (8 configs). Ensemble uses $K \in \{1, 4\}$ and $R \in \{1, 2, 4\}$ with the LLM aggregator (6 configs). Hierarchical Orchestrator uses $D \in \{0, 2, 3, 5\}$ and $T \in \{1, 2\}$, with `direct` mode for D=0 and `freeform` mode for $D \geq 2$ (8 configs). We apply a JSON-strict completeness filter that retains configurations with 228 per-question provenance JSON files; documented top-level API failures are excluded from the scored denominator. Thus the main `mini` analysis reports 22 configurations. Ensemble K=4, R=4 has 228 raw JSON files, but two top-level API failures, so it contributes 226 scored predictions. The fuller `nano` sweep, including the additional Predictor–Critic T=1 settings, is reported in the released metrics files and summary tables.

Baselines. We compare the three multi-agent architectures against each other in a controlled, like-for-like setup; standard single-pass and self-consistency baselines (zero-shot CoT (Wei et al., 2022) and self-consistency (Wang et al., 2023) with $N \in \{4, 8, 16\}$) are not included in the headline numbers. We discuss this scope choice and its implications in Section X.

Variance estimate. We report single-seed full-benchmark results and quantify uncertainty across questions using paired bootstrap intervals (Section T). Multi-seed variance remains a limitation because of the cost of full provenance-logged reruns; we discuss this in Section X and design a small targeted seed study (best configuration per architecture, 50-question subset) as future work.

T. Statistical significance tests

The main results compare the best configuration of each architecture on the same 228 questions. Because the peak accuracies differ by at most 4.0 percentage points, we treat raw-accuracy ordering cautiously and use paired tests to quantify uncertainty.

Table 19 reports paired bootstrap confidence intervals for accuracy differences and exact two-sided McNemar tests.

Table 19. Paired accuracy comparisons between the best configuration of each architecture on the same 228 questions. Bootstrap CI: paired 10 000-resample. McNemar: exact two-sided.

A	B	ΔAcc 95% CI	McNemar p
Ensemble K=1, R=2	Orchestrator D=2, T=2	$[-0.088, +0.035]$	0.49
Ensemble K=1, R=2	Pred.–Critic T=5, E=5	$[-0.039, +0.066]$	0.75
Orchestrator D=2, T=2	Pred.–Critic T=5, E=5	$[-0.018, +0.101]$	0.25

U. Cross-model full results

The complete `gpt-5.4-mini` and `gpt-5.4-nano` sweeps are written to the released metrics files. For the quick five-model Orchestrator comparison in Section A.4.5, we restrict to the selected delegation-enabled settings and require complete 228-JSON provenance. Table 20 reports the best row per model; the full file includes all 18 complete delegation-enabled Orchestrator configs.

Table 20. Best complete Orchestrator configuration per model among the selected delegation-enabled settings.

Model	Best delegation-enabled config	Correct	Acc. (%)	Cost ($/q)
`gpt-5.4-mini`	D=2, T=2 freeform	184/228	80.7	0.175
Claude Sonnet 4.5	D=2, T=2 freeform	183/228	80.3	0.912
DeepSeek v4 Flash	D=3, T=1 freeform	183/228	80.3	0.037
Gemini 2.5 Flash	D=3, T=2 freeform	163/228	71.5	0.165
`gpt-5.4-nano`	D=3, T=2 freeform	162/228	71.1	0.070

The full `nano` architecture sweep is useful as a robustness check rather than the main headline. Its best raw-accuracy configuration is Ensemble K=4, R=4 at $166/228 = 72.8\%$, while the best Orchestrator is D=3, T=2 at $162/228 = 71.1\%$. The paired accuracy difference is not statistically resolved: the bootstrap 95% CI for Ensemble minus Orchestrator is $[-4.4, +8.3]$ percentage points and McNemar's exact two-sided $p = 0.68$. The cost gap is large, however: the Orchestrator costs \$0.070/question versus \$0.899/question for the Ensemble (about 13× cheaper). Thus the nano sweep supports the cost–quality conclusion even though its raw best point differs from the `mini` sweep.

V. Metadata-based contamination-filter recall audit

The web filter emits per-page pass/fail decisions during retrieval. To audit whether it catches obvious leakage, we construct metadata-based proxy labels from page dates, result snippets, and filter metadata. This is deliberately a proxy: it catches pages that look leakage-prone from metadata, not a human judgement of whether the page truly contaminates the forecast. Across 847431 page-level filter decisions, the broad proxy identifies 141159 likely leakage pages, of which 126563 are blocked, giving proxy recall 89.7% with Wilson 95% CI $[89.5, 89.8]\%$. Under a stricter proxy definition, 68301 pages are flagged and 62876 are blocked, giving proxy recall 92.1% with CI $[91.9, 92.3]\%$.

The audit also records candidate misses: pages retained by the filter but flagged by the proxy. These are useful for filter debugging and for sampling a future human annotation set. We treat such annotation as future work: the proxy audit supports the claim that the filter catches most obvious metadata-level leakage, but it does not validate semantic precision or recall on its own.

W. Cost breakdown full results

Table 3 in the main text reports representative percentage splits between agent/model calls and web-filter calls. Figure 19 gives the full percentage view, and Figure 20 shows the same phenomenon in dollar units. The conclusion is unchanged: the web-filtering layer dominates total pipeline cost across the corrected `mini` sweep, while the Orchestrator configurations remain at the low-cost end because they retrieve less evidence and invoke fewer filtered pages.

This appendix view is useful for separating two facts that are compressed in the main-text summary table. First, the web filter is the majority of spend for every configuration. Second, the absolute dollar totals still vary substantially across architectures:

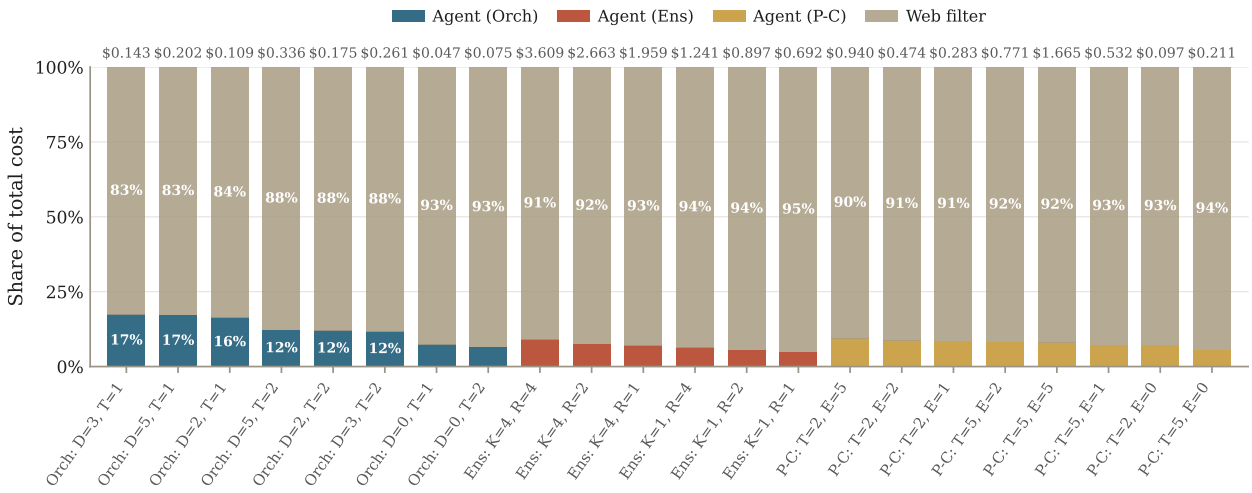


Figure 19. Full cost-share decomposition per configuration. The contamination filter accounts for 83–95% of total spend across the `mini` sweep, with the Orchestrator family at the lower end of that range.

static breadth and depth incur larger retrieval/filtering bills because their macro-schedules trigger more tool use, whereas the Orchestrator's history-conditioned routing keeps both agent-side and filter-side costs lower.

X. Limitations

Our results should be interpreted as a controlled empirical study of compute allocation for agentic forecasting, not as a complete evaluation of all possible forecasting systems or deployment settings. The main limitations concern model and architecture coverage, statistical power, contamination-filter validation, benchmark domain balance, and the extent to which API costs proxy for underlying compute.

- **Partial non-OpenAI architecture coverage.** We have complete `gpt-5.4-nano` and `gpt-5.4-mini` architecture sweeps (Orchestrator, Ensemble, Predictor–Critic), plus complete subsets of Orchestrator runs for Claude Sonnet 4.5, DeepSeek v4 Flash, and Gemini 2.5 Flash. The five-model Orchestrator comparison in Section A.4.5 should therefore be read as a complete-config snapshot for selected delegation-enabled settings, not as a full-factorial model-family sweep; a full architecture $\times$ parameter $\times$ model sweep is left for future work.

- **Standard test-time-compute baselines deferred.** We plan to add zero-shot chain-of-thought and self-consistency ($N \in \{4, 8, 16\}$) baselines on the 228-question benchmark in future work. The within-paper architecture comparisons are unaffected by their absence.

- **Single seed per configuration.** We rely on N=228 questions for variance reduction rather than seed averaging. Multi-seed estimation on targeted subsets remains future work.

- **Statistical resolution.** At N=228 and the observed inter-architecture gap ($\leq$ 4.0 percentage points in the `mini` sweep), the McNemar test is not powered to resolve accuracy differences among best configurations (Section T); the robust claim is cost–quality dominance, not a statistically decisive raw accuracy ordering.

- **Filter-recall validation.** The metadata-based proxy recall audit is a scalable sanity check, not human ground truth. A targeted human annotation study of candidate misses and blocked pages remains necessary to validate the actual contamination-filter recall and precision.

- **Domain coverage.** Topic distribution is set by ForecastBench; rare domains (Science/Climate n=3, Entertainment n=5) are underrepresented.

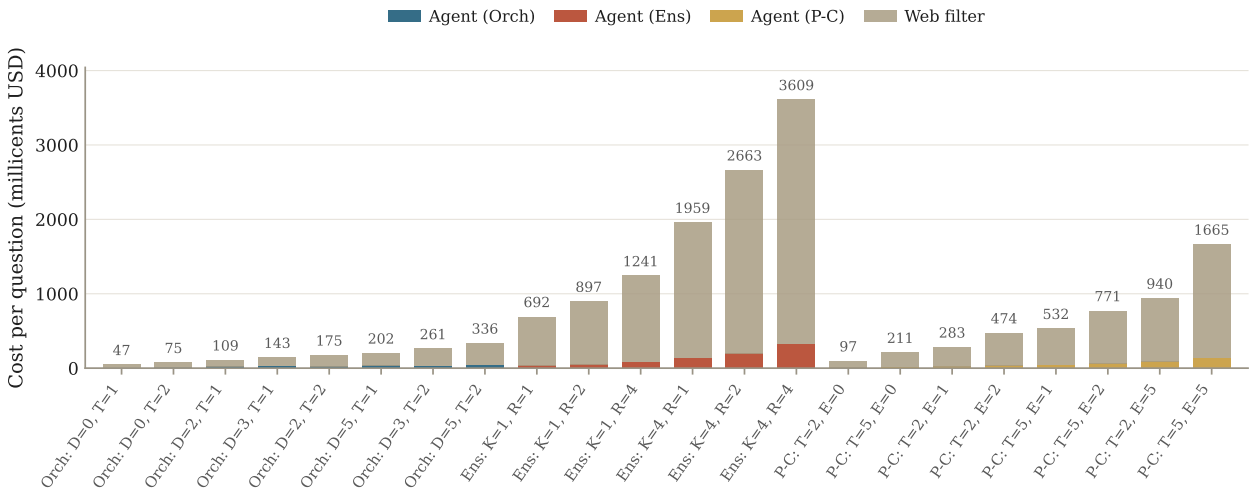


Figure 20. Cost decomposition per configuration. The contamination filter accounts for 83–95% of total spend across all `mini` configurations (median 91.8%); the Orchestrator family sits at the lower end of this range because its retrieval budget is history-conditioned.

- **Tool surface and delegation mode.** The freeform orchestrator defines sub-agent roles at runtime; a predefined variant (fixed Research / Data / Analysis sub-agents) was implemented but not swept. Whether different fixed-role designs or hybrid predefined/freeform schedules would change the cost–accuracy tradeoff is an open question.
- **Cost-as-proxy.** USD pricing reflects current OpenAI rates; we therefore report token- and call-based proxies alongside USD throughout (Section A.4).