WHEN IS MORE REASONING WORTH IT? FALLBACK-AWARE REASONING-STRATEGY ROUTING UNDER HARD TOKEN BUDGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning strategies have heterogeneous, input-dependent token costs. Standard test-time allocation asks where compute is useful; we study the distinct problem of hard per-query output-token compliance, where realized use above a requested threshold is a violation. We separate this risk into violations when a strategy is predicted feasible and violations after no strategy is predicted feasible, then use a frozen router combining strategy-specific token prediction, estimator-fit accuracy priors, conservative resource corrections, and minimum-predicted-token fallback. In a sealed true-online evaluation of Qwen2.5-3B-Instruct and Phi-3.5-mini-instruct across four reasoning benchmarks, 2,400 routed executions precede a 4,000-execution exhaustive matrix. From low to high budgets, macro violation falls from 0.258 to 0.033 for Qwen and from 0.225 to 0.038 for Phi, while fallback falls from 0.688 to 0.123 and from 0.750 to 0.048, respectively. Yet conservatism is not uniformly beneficial: against the frozen point router, Phi's low-budget MATH-500 violation increases by 0.25 (95% bootstrap CI $[0.15, 0.35]$; Holm-adjusted $p = 0.00024$), and low-budget fallback carries substantial model-dependent accuracy costs. Hard-budget reasoning therefore requires controlling both uncertainty within the feasible branch and the behavior of the system when predicted feasibility disappears.

1 INTRODUCTION

Test-time scaling and routing decide where finite inference compute should be spent: how long to reason, which model to invoke, or which procedure to apply. Chain-of-Thought prompting, repeated critique, and iterative revision can help on difficult inputs, but they consume heterogeneous and input-dependent resources (Wei et al., 2022; Madaan et al., 2023; Snell et al., 2025). Existing work provides increasingly capable mechanisms for allocating compute, cost, model capacity, or reasoning budget (Ong et al., 2024; Chen et al., 2023; 2026; Fernandez et al., 2026; Kapoor et al., 2026).

Hard per-query compliance poses a different objective. A user supplies an output-token budget B; realized generation length is uncertain; and exceeding B is itself a failure, even when average consumption is acceptable. In short, **compute allocation is not the same as hard-budget compliance**. A router must reason about the distributional error between predicted and realized token use, not merely rank strategies by expected utility or cost.

Uncertainty creates a second difficulty: conservative estimates can leave no strategy predicted feasible. The system must still define what happens. Its fallback choice then creates a separate risk channel, with its own frequency and conditional violation rate. Treating fallback as an implementation detail can therefore hide the dominant source of end-to-end failure at tight budgets.

This leads to the paper's central observation. Greater conservatism may improve budget behavior among strategies declared feasible while increasing the chance that the feasible set is empty. If fallback is risky, the second effect can offset or reverse the first. We make this trade-off explicit, build it into a frozen reasoning-strategy router, and evaluate it separately from conventional accuracy-resource allocation.

Our three scientific contributions are:

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

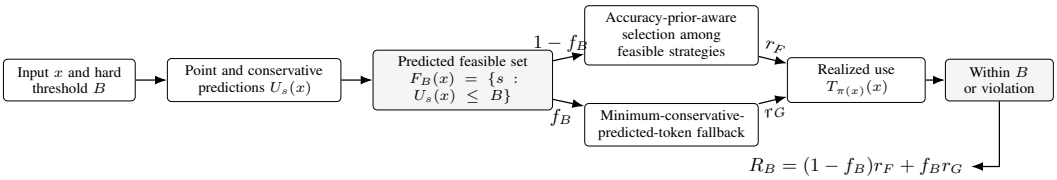


Figure 1: Fallback-aware routing under a hard compliance threshold. Greater feasible-branch protection need not lower total violation if conservatism increases fallback exposure. The decision uses frozen input-derived features, resource predictions, and estimator-fit accuracy priors. No current-query gold label, confirmatory or exhaustive outcome, or hindsight-oracle information enters routing.

1. We formulate reasoning-strategy routing under explicit per-query output-token budgets and separate feasible-branch and fallback-branch risk.
2. We develop a frozen fallback-aware router combining strategy-specific token prediction, conservative resource estimates, strategy accuracy priors, and minimum-conservative-predicted-token fallback.
3. Across two model families and four benchmarks, we study whether conservatism improves feasible-branch compliance yet induces model-dependent end-to-end trade-offs through fallback.

As an evaluation-rigor measure rather than algorithmic novelty, we freeze all development decisions before confirmatory access and execute true-online routing before exhaustive confirmatory measurement. Figure 1 summarizes the resulting fallback-aware decision path.

2 Related Work

Test-time compute scaling. Chain-of-Thought prompting demonstrates that inference procedure can alter reasoning performance without parameter updates (Wei et al., 2022). Self-Refine uses model-generated feedback and revision (Madaan et al., 2023). The present work treats such procedures as fixed strategy candidates and asks which one to invoke under an explicit output-token limit. Snell et al. (Snell et al., 2025) study optimal allocation of test-time compute for reasoning, and Yue et al. (Yue et al., 2025) study inference scaling for long-context retrieval-augmented generation. These works motivate adaptive allocation; our outcome of interest is instead whether uncertain realized use respects a hard per-query limit.

Adaptive inference and routing. RouteLLM learns routing policies from preference data (Ong et al., 2024), while FrugalGPT studies cascades across model services (Chen et al., 2023). CONCUR supports continual constrained and unconstrained routing with separate strategy predictors (Chen et al., 2026); RADAR routes model-budget pairs using estimated query difficulty and model ability (Fernandez et al., 2026); and TRIM intervenes stepwise by routing selected reasoning steps to stronger models (Kapoor et al., 2026). We do not claim budgeted routing itself as new. Our focus is query-level selection among heterogeneous reasoning strategies, uncertainty in their realized token cost, and the explicit event that no strategy is predicted feasible.

Budget-aware reasoning. Plan-and-Budget allocates test-time budget across decomposed subquestions (Lin et al., 2026), whereas we predict the realized resource use of a single query-level strategy choice. BOLT studies budget-aware routing for constrained multimodal robot question answering (Ni et al., 2026). It is closely related in resource-aware motivation but differs in domain and routing mechanism. TALE estimates and communicates reasoning-token budgets according to problem complexity (Han et al., 2025). We instead select among heterogeneous reasoning strategies, predict each candidate's realized output-token requirement before execution, count threshold exceedance as a compliance failure, and analyze the explicit fallback event when no candidate is predicted feasible. Our emphasis is measured hard-budget compliance and fallback risk, not merely expected or allocated resource use.

Resource and risk calibration. Prompt Risk Control provides a statistical framework for controlling deployment risk (Zollo et al., 2024), while SAFER studies risk-constrained sample-then-filter generation (Wang et al., 2026). These methods motivate careful separation between empirical conservatism and formal guarantees. Our operational router does not provide their style of finite-sample risk guarantee: its risk-calibration data are used both to compare and select among estimator families. A pre-Stage-D joint conformal alternative is retained only as a development diagnostic (Appendix F).

3 Problem Formulation

Let x be an input, $\mathcal{S}$ the frozen strategy set, B a per-query hard compliance threshold, $T_s(x)$ the random generated-token usage of strategy s, and $A_s(x)$ its correctness or utility. At the conceptual population level, randomness can represent stochastic decoding or deployment variation, and an ideal conditional-risk router would solve

$$\max_{s\in\mathcal{S}} \mathbb{E}[A_s(X) \mid X = x] \quad \text{subject to} \quad \Pr(T_s(X) > B \mid X = x) \le \delta. \tag{1}$$

The frozen experiment instead uses deterministic decoding (fixed revisions, prompts, and configuration; temperature 0; sampling disabled). Its implemented router does not estimate the conditional probability in Eq. (1); it uses a learned pre-execution conservative token estimate $U_s(x)$ and estimator-fit-only expected-accuracy priors, without a conditional-risk guarantee.

The predicted feasible set is

$$F_B(x) = \{s \in \mathcal{S} : U_s(x) \le B\}. \tag{2}$$

If this set is non-empty, the router chooses the candidate with the highest frozen accuracy prior, then lower conservative prediction, then frozen candidate order. Otherwise it invokes the explicit fallback defined below.

4 Fallback-Aware Budget Risk

Define the empty-feasible-set event $E_B = \{F_B(X) = \emptyset\}$ and budget violation event $V_B = \{T_{\pi(X)}(X) > B\}$. Let

$$f_B = \Pr(E_B), \qquad r_F = \Pr(V_B \mid E_B^c), \qquad r_G = \Pr(V_B \mid E_B). \tag{3}$$

The law of total probability gives

$$R_B = \Pr(V_B) = (1 - f_B)r_F + f_B\, r_G = r_F + f_B(r_G - r_F). \tag{4}$$

For a point router p and a more conservative router c, it is possible that $r_{F,c} < r_{F,p}$ but $R_c > R_p$. By subtracting the two forms of Eq. (4), this occurs exactly when

$$f_c(r_{G,c} - r_{F,c}) - f_p(r_{G,p} - r_{F,p}) > r_{F,p} - r_{F,c}. \tag{5}$$

Equation (5) is an algebraic proposition, not a novel theorem. Its scientific role is to connect feasible-branch calibration to the empirically important exposure induced by fallback.

When $F_B(x)$ is empty, the frozen fallback is

$$g(x) = \arg\min_{s\in\mathcal{S}} U_s(x), \tag{6}$$

with ties broken by frozen candidate order. This rule minimizes the conservative predicted token requirement and therefore any nondecreasing function of the predicted excess $\max\{U_s(x) - B, 0\}$. It is a surrogate-optimality statement only; without an additional stochastic-order assumption it does not imply minimum true violation probability.

5 Budget-Aware Reasoning Router

Strategies. The fixed candidates are Direct, Chain-of-Thought, Few-Shot, Self-Refine, and Critic-Refine, in that tie-breaking order. Direct, Chain-of-Thought, and Few-Shot use one generation call; Self-Refine uses an initial response and one revision; Critic-Refine adds a distinct critique before revision. Demonstrations are fixed per dataset and disjoint from all experimental roles.

Resource prediction and priors. Strategy-specific nonnegative linear point predictors are fitted exclusively on the estimator-fit role. Candidate conservative corrections are evaluated on the risk-calibration role. Complexity-stratified estimator-fit success rates provide strategy accuracy priors. For each model–dataset pair, the frozen estimator selection evaluates each candidate on 200 successful strategy–example executions (40 risk-calibration examples times five strategies). Each execution is one Bernoulli observation of whether measured output tokens exceed that candidate's prediction, and the mean prediction is averaged over the same 200 executions. Selection chooses the lowest-mean-predicted-token candidate whose upper 95% Wilson bound on this development violation rate is at most 0.15; if none qualifies, it uses the pooled 0.95 residual candidate. Analysis requires all 200 records to be successful rather than dropping failures. Because the same risk-calibration observations compare and select estimator families, no pristine post-selection conformal guarantee is claimed.

Routing. Among predicted-feasible candidates, the router orders by highest frozen accuracy prior, lower conservative token prediction, and candidate order. If no candidate is predicted feasible, it chooses the minimum conservative predicted token requirement, again using candidate order only for ties. Decision inputs exclude measured tokens, correctness, evaluator output, exhaustive outcomes, confirmatory outcomes, and hindsight oracles.

Budgets. Low, medium, and high budgets are model- and dataset-specific nearest-rank p25, p50, and p75 quantiles of the 200 measured risk-calibration output-token counts pooled over the five strategies and 40 examples. For sorted values $t_{(1)} \leq \cdots \leq t_{(n)}$, the frozen convention returns $t_{(\max(1,\lceil qn \rceil))}$ for quantile q. Each B is treated as a hard compliance threshold: measured generated-token use greater than B is a violation. Evaluation does not mechanically truncate generation at B; it retains the larger frozen generation safety caps described below. Using generation safety caps larger than B allows overshoot relative to the routing threshold to be observed rather than mechanically set to zero. Across D1/D3, 133 of 8,842 calls (1.50%) nevertheless reached those safety caps (Appendix I). Because every safety cap exceeds the corresponding low, medium, and high routing thresholds, cap-hit calls remain unambiguous budget violations, although their recorded token lengths are right-censored and their answer quality may reflect truncation. Cross-model comparisons concern qualitative replication rather than identical absolute token limits. The exact frozen thresholds for every model–dataset pair are reported in Table 5.

6 Experimental Setup

Models. The primary model is Qwen/Qwen2.5-3B-Instruct at revision `aa8e72537993ba99e69dfaafa59ed015b17504d1`. The second family is microsoft/Phi-3.5-mini-instruct with both model and tokenizer fixed at revision `2fe192450127e6a83f7441aef6e3ca586c338b77`. Both use BF16 CUDA execution without quantization. Generation is deterministic with temperature 0 and sampling disabled. Per-call generation safety caps are 576 tokens for ARC-Challenge, 1,024 for GSM8K and MMLU-Pro, and 1,536 for MATH-500; these caps are distinct from the measured routing budgets. Across the 8,842 recorded D1/D3 model calls, explicit finish metadata identifies 133 safety-cap terminations (1.50%); Appendix I gives the breakdown. Generation configuration, completion contracts, and strategy prompts remain frozen from development.

Datasets and roles. We use GSM8K (Cobbe et al., 2021), ARC-Challenge (Clark et al., 2018), MMLU-Pro (Wang et al., 2024), and the numeric-evaluable MATH-500 subset, derived from the MATH benchmark (Hendrycks et al., 2021). Each dataset contributes 60 estimator-fit, 40 risk-calibration, and 100 sealed-confirmatory identities, plus three fixed demonstrations. Selection uses deterministic SHA-256 ranking after Unicode NFC and collapsed-whitespace normalization. Answer-free role manifests were frozen before development inference. Identity and normalized-content audits found no V1/V2 overlap or cross-role duplication. Estimator-fit data fit point models; risk-calibration data select conservative corrections and construct budgets; sealed-confirmatory data are reserved for the final frozen evaluation. The resource-aware fallback amendment and rejection of the joint conformal alternative both occurred before sealed access. Confirmatory identities were unsealed only after the router, fallback, budgets, analysis families, and execution order were frozen.

Execution order and latency. Stage D is pre-specified and frozen before sealed-confirmatory access to run true-online routed evaluation first, immutably archive it, then run the exhaustive five-strategy confirmatory matrix and derive offline analyses. The router remains frozen throughout. Latency uses `time.perf_counter`, synchronizes CUDA immediately before and after each generation boundary, excludes model loading, includes all scientific examples after one excluded synthetic warmup per model/strategy call shape, and reports mean, median, p25, p75, and p95. Routing, prompt construction, generation, and evaluation components are recorded separately.

Statistics. Pre-specified confirmatory comparisons use exact two-sided McNemar tests for paired accuracy and violation outcomes, two-sided Wilcoxon signed-rank tests with Pratt zero handling for paired resources, Wilson violation intervals, and Holm correction within each pre-specified metric family. The D1 families are accuracy versus Always-Direct and tokens versus Always-Direct (24 hypotheses each: two models, four datasets, and three tiers). Offline families are V2 versus V1 violation and V2 versus the estimator-fit best fixed strategy in accuracy (24 each), plus adjacent-tier ordered accuracy and token contrasts (16 each: two models, four datasets, and two adjacent-tier contrasts). Bootstrap intervals use 10,000 replicates and seed 20260908. Macro results are unweighted means of dataset-level metrics.

7 Confirmatory Evaluation

RQ1: Does conservative resource prediction improve hard-budget compliance? The answer is conditional. Relative to the frozen point router, V2 reduces violations significantly after Holm correction in Qwen ARC at medium (-0.21, 95% CI $[-0.30, -0.12]$, $p_{\mathrm{H}} = 0.00041$) and high (-0.15, $[-0.22, -0.08]$, $p_{\mathrm{H}} = 0.00122$), and Qwen MMLU-Pro at medium (-0.11, $[-0.17, -0.05]$, $p_{\mathrm{H}} = 0.0176$) and high (-0.13, $[-0.20, -0.07]$, $p_{\mathrm{H}} = 0.00464$). Phi MMLU-Pro likewise improves at medium (-0.29, $[-0.38, -0.20]$, $p_{\mathrm{H}} < 10^{-6}$) and high (-0.28, $[-0.38, -0.18]$, $p_{\mathrm{H}} < 10^{-5}$). The counterexample is Phi MATH-500 at low budget, where violation worsens by 0.25 ($[0.15, 0.35]$, $p_{\mathrm{H}} = 0.00024$). Thus larger budgets yield lower absolute violation, but the method-specific comparison is neither trivial nor universal.

RQ2: When does fallback dominate end-to-end risk? Fallback is frequent at tight budgets: macro f_B is 0.688 for Qwen and 0.750 for Phi at low budget, falling to 0.123 and 0.048 at high budget. MATH-500 is the sharpest failure mode: both models fall back on every low-budget example, with observed violation 0.53 for Qwen and 0.62 for Phi. The persisted cell-level decomposition reconstructs every observed R_B exactly. Fallback is therefore part of the end-to-end estimand, not a censored implementation condition.

Table 2 exposes the macro decomposition behind these rates. At low budget, fallback accounts for 102 of Qwen's 103 violations and all 90 Phi violations; as budgets loosen, both fallback exposure and its contribution fall.

RQ3: Are accuracy-resource-fallback trade-offs consistent across models? Qwen accuracy is comparatively stable across low/medium/high tiers (0.600/0.583/0.595) while mean output tokens rise (157.9/196.5/265.5). Phi shows a stronger accuracy-resource progression (0.473/0.540/0.593) as tokens rise (99.7/109.2/144.0). Absolute budgets are model- and dataset-specific, so this is qualitative replication rather than an identical-budget comparison. Strategy choice is concentrated: Direct and Few-Shot account for all low- and medium-tier decisions; Critic-Refine is never selected.

RQ4: How does the router compare with fixed strategies and frozen baselines? The exhaustive matrix supports five always-fixed policies, a best fixed policy selected using estimator-fit accuracy then resource use, frozen V1 routing, five frozen ablations, and an analysis-only oracle. V2 does not broadly exceed the best fixed baseline in accuracy. For Phi GSM8K it is significantly worse at low (-0.51, $[-0.61, -0.40]$, $p_{\mathrm{H}} < 10^{-11}$) and medium budget ($-0.28$, $[-0.37, -0.19]$, $p_{\mathrm{H}} < 10^{-5}$). These results locate V2's value in conditional compliance-resource trade-offs rather than universal accuracy dominance.

Table 1: Sealed true-online macro results. Routing overhead is separate from generation.

Model	Tier	Acc.	Viol.	Fallback	Mean tok.	Route ms
Qwen	Low	.600	.258	.688	157.9	1.885
Qwen	Medium	.583	.125	.438	196.5	1.881
Qwen	High	.595	.033	.123	265.5	1.843
Phi	Low	.473	.225	.750	99.7	2.109
Phi	Medium	.540	.078	.435	109.2	2.063
Phi	High	.593	.038	.048	144.0	2.055

Table 2: True-online macro fallback decomposition. Rates use exact pooled branch counts over 400 examples per row; $\widehat{R}_B$ is reconstructed from the three preceding columns and R_B is observed.

Model	Tier	f_B	r_F	r_G	$\widehat{R}_B$	R_B
Qwen	Low	.688	.008	.371	.258	.258
Qwen	Medium	.438	.080	.183	.125	.125
Qwen	High	.123	.026	.082	.033	.033
Phi	Low	.750	.000	.300	.225	.225
Phi	Medium	.435	.027	.144	.078	.078
Phi	High	.048	.034	.105	.038	.038

RQ5: What are the runtime and routing-overhead costs? Mean routing overhead is 1.84–1.88 ms for Qwen and 2.06–2.11 ms for Phi, versus mean generation latency of 9.61–16.60 s and 5.32–7.15 s across tiers. These are reproducible local-hardware measurements, not production latency.

Table 1 reports the primary true-online macro results, while Table 3 summarizes selected frozen V2-versus-V1 violation effects. Complete per-dataset results, strategy selection, fixed-strategy baselines, resource accounting, and frozen ablations appear in the appendix.

8 Analysis and Ablations

The frozen analysis compares the conservative V2 router with the V1-style point router, five always-fixed strategies, and the best fixed strategy selected by frozen estimator-fit accuracy/resource criteria. Frozen ablations remove complexity features, accuracy priors, or conservative resource correction and compare pooled with strategy-conditioned corrections. Hindsight oracle results are analysis-only accuracy upper bounds and never influence selection. Removing complexity, accuracy priors, or conservative correction changes selection and fallback exposure in model-specific ways (Appendix J); none dominates the frozen policy across all cells.

9 Discussion

Conservative calibration is insufficient by itself. Tightening $U_s(x)$ can protect the feasible branch while moving examples into fallback; whether total risk improves then depends on both f_B and r_G. This explains why locally cleaner resource estimates coexist with the Phi MATH-500 reversal and the large low-budget accuracy cost on Phi GSM8K. The development-only joint conformal diagnostic reinforces the system lesson: a cleaner simultaneous resource statement can still induce pathological fallback. It was rejected before sealed access and is not the operational method.

The two models also show why fallback cannot be treated as a model-independent constant. Their frozen predictors, priors, budgets, and generated-length distributions lead to different feasible sets and fallback choices. Hard-budget systems should report feasible-branch and fallback-branch behavior separately, validate fallback as part of the policy, and avoid presenting average allocation as evidence of per-query compliance.

Table 3: Selected frozen V2-minus-V1 violation effects. CIs are paired-bootstrap 95%; p_{H} is Holm-adjusted within the frozen violation family.

Model	Dataset	Tier	Effect [95% CI]	p_{H}	Direction
Qwen	ARC	Med.	$-.21[-.30, -.12]$	.00041	improves
Qwen	ARC	High	$-.15[-.22, -.08]$	.00122	improves
Qwen	MMLU-Pro	Med.	$-.11[-.17, -.05]$	.0176	improves
Qwen	MMLU-Pro	High	$-.13[-.20, -.07]$	.00464	improves
Phi	MMLU-Pro	Med.	$-.29[-.38, -.20]$	$< 10^{-6}$	improves
Phi	MMLU-Pro	High	$-.28[-.38, -.18]$	$< 10^{-5}$	improves
Phi	MATH-500	Low	$+.25[+.15, +.35]$	.00024	worsens

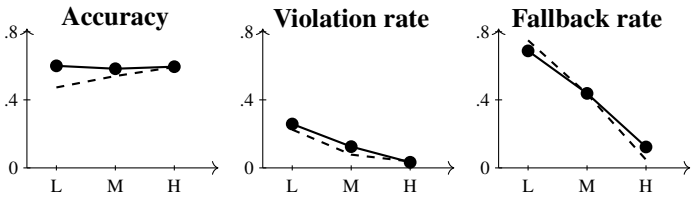


Figure 2: Sealed-confirmatory true-online macro behavior. Solid circles denote Qwen and dashed squares Phi; L/M/H denote budget tiers.

10 Limitations

The study covers two small instruction-model families and four reasoning benchmarks. It studies generated-token budgets rather than monetary cost, energy, memory, or a comprehensive resource objective. The candidate set is restricted to five fixed reasoning modes, and confirmatory routing is concentrated in a subset of them; Critic-Refine is never selected. Fallback can carry a substantial, model-dependent accuracy cost. The selected operational estimator has no universal or post-selection conformal guarantee. Latency depends on the frozen local hardware and measurement protocol and should not be interpreted as production latency. Exact uninterrupted wall time is unavailable because of documented pauses and laptop shutdowns. Conclusions cannot be generalized beyond the evaluated models, datasets, prompts, or strategy implementations. Development-stage generation safety caps right-censored 101 of 6,400 calls used in fitting and calibration. A post-hoc sensitivity audit left all eight frozen budget triplets invariant, but three model–dataset estimator or margin cells were not provably invariant to arbitrary uncensoring; exclusion-refitting the point models changed selected-strategy identity in 58 of 2,400 tier-specific confirmatory routing decisions and fallback status in 32 of 2,400 (Appendix I). Robustness is therefore not universal: for example, on Qwen MATH-500 at medium budget, violation is 0.27 for the V1 point router and 0.26 for frozen V2, but rises to 0.28 under the point-model exclusion sensitivity.

11 Conclusion

Hard-budget reasoning is not solved by allocating more or less compute on average. The frozen two-model evaluation shows that conservative routing can improve compliance in important cells, but can also fail when it expands exposure to a poor fallback. The design lesson is direct: hard-budget routing must reason about uncertainty among feasible choices and about what happens when predicted feasibility disappears.

12 Reproducibility Statement

The protocol pins model and tokenizer revisions, dataset revisions, normalized answer-free role manifests, prompts, strategies, evaluator behavior, estimator candidates, selection criteria, budgets, statistics, and execution order before confirmatory access. Canonical SHA-256 fingerprints identify immutable configuration and component snapshots. Append-only JSONL records use complete

deterministic execution identities and resume without inferring completion from example ID alone. Regeneration checks, role-disjointness audits, fixed seed 20260908, explicit failure records, and pre-confirmatory router fingerprints support independent audit. No identifying local filesystem paths are required to reproduce the protocol.

13 AI Assistance Disclosure

Generative AI tools assisted with literature discovery, research ideation and methodology feedback, implementation and experiment-execution support, code generation and editing, manuscript structure, drafting and polishing, and the formulation and checking of mathematical claims. The human author made all final scientific and protocol decisions; sealed-data access and experiment authorization required explicit human approval. Generated code was tested and validated, and the author reviewed cited references and quantitative claims against source material and frozen evidence. The author takes responsibility for the final submission. No identifying conversation data are included.

References

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

Peter Baile Chen, Weiyue Li, Dan Roth, Mike Cafarella, Samuel Madden, and Jacob Andreas. CONCUR: A framework for continual constrained and unconstrained routing. In *International Conference on Learning Representations*, 2026.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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

Nigel Steven Fernandez, Branislav Kveton, Ryan Rossi, Andrew Lan, and Jack Wang. RADAR: Reasoning-ability and difficulty-aware routing for reasoning LLMs. In *International Conference on Learning Representations*, 2026.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware LLM reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24842–24855, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.1274. URL https://aclanthology.org/2025.findings-acl.1274/.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Vansh Kapoor, Aman Gupta, Hao Chen, Anurag Beniwal, Jing Huang, and Aviral Kumar. TRIM: Hybrid inference via targeted stepwise routing in multi-step reasoning tasks. In *International Conference on Learning Representations*, 2026.

Junhong Lin, Xinyue Zeng, Jie Zhu, Song Wang, Julian Shun, Jun Wu, and Dawei Zhou. Plan and budget: Effective and efficient test-time scaling on reasoning large language models. In *International Conference on Learning Representations*, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2023.

Tengjun Ni, Xin Yuan, Shenghong Li, Kai Wu, Ren Liu, Wei Ni, and Wenjie Zhang. BOLT: Decision-aligned distillation and budget-aware routing for constrained multimodal QA on robots. In *International Conference on Learning Representations*, 2026.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025.

Qingni Wang, Yue Fan, and Xin Wang. SAFER: Risk-constrained sample-then-filter in large language models. In *International Conference on Learning Representations*, 2026.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. Inference scaling for long-context retrieval augmented generation. In *International Conference on Learning Representations*, 2025.

Thomas Zollo, Todd Morrill, Zhun Deng, Jake Snell, Toniann Pitassi, and Richard Zemel. Prompt risk control: A rigorous framework for responsible deployment of large language models. In *International Conference on Learning Representations*, 2024.

A Frozen Protocol Summary and Fingerprints

The frozen router fingerprint is `633cef5c419c165b75b28a1383c5a753 4590aa9b94ff30aba977102829aff7af`; the resource-aware fallback fingerprint is `a3ec7e3dbe191964de55e4dc65ab1ab29ac3bf50188c86dd042bf8b6fa9a0137`. Five strategies and candidate order were fixed before sealed access. Failures remain explicit records; completed execution identities are immutable under append-only resume. True-online routing produced 2,400 successful records before the 4,000-record exhaustive matrix. The joint-conformal candidate was rejected before unsealing and never became operational.

B Dataset Manifests and Split Construction

The role order is demonstrations, estimator-fit, risk-calibration, then sealed confirmatory. Selection uses SHA-256 ranking with seed 20260908 after excluding all V1 identities and previously assigned V2 roles. Manifests contain no answers. Each dataset contributes three demonstrations, 60 estimator-fit examples, 40 risk-calibration examples, and 100 sealed-confirmatory examples. Identity and normalized-content audits establish role disjointness and V1 exclusion.

Table 4: Pinned public dataset sources. Revisions are immutable repository commits.

Dataset	Configuration	Role source split(s)	Revision
GSM8K	`main`	train / test	`740312add88f781978c0658806c59bc2815b9866`
ARC-Challenge	`ARC-Challenge`	train / test	`210d026faf9955653af8916fad021475a3f00453`
MMLU-Pro	`default`	validation / test	`b189ec765aa7ed75c8acfea42df31fdae71f97be`
MATH-500	`default`	test	`6e4ed1a2a79af7d8630a6b768ec859cb5af4d3be`

MMLU-Pro retains chemistry, computer science, engineering, mathematics, and physics. The numeric-evaluable MATH-500 pool contains 489 rows whose gold answer is deterministically accepted by the frozen numeric evaluator before sampling; after excluding 103 V1 identities, V2 assigns 3 demonstrations, 60 estimator-fit, 40 risk-calibration, and 100 sealed-confirmatory identities, leaving 183 eligible rows unused. Text normalization uses Unicode NFC and collapsed whitespace.

C STRATEGY PROMPTS

All strategies receive only task-visible question text and choices. Direct asks for the required final-answer field; Chain-of-Thought adds an instruction to reason step by step; Few-Shot prepends the three frozen demonstrations. Self-Refine generates an initial response and then requests a corrected revision. Critic-Refine generates an initial response, requests a separate critique, and revises using both. Answer calls terminate with `Final answer: <value>` followed by `END_RESPONSE`; critiques terminate with `END_CRITIQUE`. Prompts never expose gold labels or evaluator outcomes. Full literal prompt and coefficient payloads are not attached to this PDF-only review submission; their frozen identities remain bound by the protocol and analysis fingerprints below.

D PREDICTOR AND CALIBRATION DETAILS

For strategy s, the point model is $\widehat{T}_s(x) = \max(0, \beta_{0s} + \beta_{ws} w(x) + \beta_{cs} c(x))$, with nonnegative coefficients fitted on estimator-fit data only. $w(x)$ is word count and $c(x)$ is the sum of seven deterministic indicators computed from the NFC- and whitespace-normalized question: at least 35 words, at least three sentence segments, at least three numeric expressions, at least two arithmetic or comparison operators, an equation cue, at least two frozen multi-step lexical cues, and at least four distinct normalized numbers. Sentence segments split on period, exclamation, or question-mark runs; operator counts combine symbolic addition, subtraction, multiplication, division, power, and comparison forms with the frozen lexical equivalents. An equation cue is an equals sign or the phrase "equals", "is equal to", or "solve for". The multi-step lexicon is {then, after, before, remaining, altogether, each, per, respectively, first, second, finally, increase, decrease, more than, less than, twice, three times, difference, total, remainder}. There is no dataset-specific transformation. Candidate upper estimates add pooled Q90/Q95, strategy-conditioned Q90, or split-conformal residual margins. Risk-calibration data determine margins and budgets, never point coefficients or priors. Selection chooses the least conservative candidate whose development Wilson upper bound is at most 0.15, falling back to pooled Q95 if none qualifies. Table 5 provides the frozen selected estimators, exact budgets, and compact margins.

Table 5: Frozen selected conservative estimators, exact token budgets (L/M/H), and strategy margins in candidate order D/C/F/S/R (Direct, Chain-of-Thought, Few-Shot, Self-Refine, Critic-Refine). Margins are rounded to three decimals for display; the exact payloads are fingerprinted below.

Model	Dataset	Estimator	L	M	H	D	C	F	S	R
Qwen	GSM8K	strategy Q90	169	250	442	114.523	119.111	114.710	257.090	245.626
Qwen	ARC	strategy Q90	127	229	328	63.139	109.588	0.000	99.538	217.156
Qwen	MMLU-Pro	strategy Q90	267	462	856	266.492	240.120	224.798	279.554	504.212
Qwen	MATH-500	pooled Q90	301	574	846	400.452	400.452	400.452	400.452	400.452
Phi	GSM8K	pooled Q90	115	218	356	116.506	116.506	116.506	116.506	116.506
Phi	ARC	strategy Q90	13	199	326	0.991	82.533	0.000	98.908	282.879
Phi	MMLU-Pro	pooled Q90	19	362	617	393.151	393.151	393.151	393.151	393.151
Phi	MATH-500	strategy Q90	192	383	689	160.920	429.199	545.915	516.065	1025.782

The canonical exact Qwen Stage-B and Phi Stage-C analysis payload fingerprints are respectively `b590ccee38a6bf27625e79980208a32fd43a2d718e1879ef4cc752cc81dd1a84` and `d01d3d7388d369c1eab19350c6d8e30502a3bf8a21fcb115d29ff2352041369a`.

E ADDITIONAL DEVELOPMENT RESULTS

All material in this section is **DEVELOPMENT / NON-CONFIRMATORY**. Development data selected model-specific estimators and exposed the importance of resource-aware fallback. These observations motivated the frozen amendment but are not mixed with confirmatory estimates or tests.

F JOINT-CONFORMAL DEVELOPMENT DIAGNOSTIC

The rejected alternative defines the per-example score $R_i = \max_s[T_{i,s} - \widehat{T}_s(x_i)]$ and adds its fixed $\alpha = 0.10$ finite-sample upper quantile to every strategy prediction. Under standard exchangeability

assumptions and with pre-specified point predictors and score, this provides simultaneous marginal coverage for a fresh example. It does not control fallback-branch risk or imply end-to-end 90% compliance. Development leave-one-out diagnostics produced pathological fallback and model-dependent accuracy degradation, so human review rejected it for operational Stage D use. It is not presented as a novel conformal method, and the frozen V2 router is unchanged.

G Statistical Procedures

Paired binary outcomes use exact two-sided McNemar tests. Resource differences use two-sided Wilcoxon signed-rank tests with Pratt zero handling. Violation rates use Wilson 95% intervals. Paired percentile-bootstrap intervals use 10,000 replicates and seed 20260908. Holm correction is applied within the six families and hypothesis counts specified in Experimental Setup. No scientific execution failed.

H Additional Confirmatory Tables

The tables below are deterministic derivations of immutable D1/D3 records. The oracle is a non-deployable analysis-only accuracy upper bound over the five measured D3 outcomes for each example. It chooses any correct strategy if one exists, otherwise all strategies are eligible; within that correctness tier it minimizes realized output tokens, breaking ties by frozen candidate order. It does not impose realized-token compliance, so its violation rate can exceed the router's. D1 values are measured from independently executed routed calls, whereas Table 8 joins frozen selections to separately executed D3 outcomes. Their execution keys and derived seeds differ; Few-Shot also uses the seed to order demonstrations, and the CUDA runtime is not claimed to be bitwise reproducible, so the two measurements need not be numerically identical.

Table 6: Complete sealed true-online model–dataset–tier results. Tok. is mean output tokens.

Model	Dataset	Tier	Acc.	Viol.	Fallback	Tok.
Qwen	GSM8K	Low	.83	.40	1.00	155.8
Qwen	GSM8K	Med.	.83	.06	.81	155.8
Qwen	GSM8K	High	.82	.01	.00	174.9
Qwen	ARC	Low	.76	.01	.00	9.9
Qwen	ARC	Med.	.70	.09	.00	129.5
Qwen	ARC	High	.74	.01	.00	151.6
Qwen	MMLU-Pro	Low	.27	.09	.75	62.6
Qwen	MMLU-Pro	Med.	.28	.09	.00	90.7
Qwen	MMLU-Pro	High	.30	.06	.00	325.5
Qwen	MATH-500	Low	.54	.53	1.00	403.3
Qwen	MATH-500	Med.	.52	.26	.94	410.1
Qwen	MATH-500	High	.52	.05	.49	409.8
Phi	GSM8K	Low	.44	.20	1.00	51.5
Phi	GSM8K	Med.	.61	.05	.00	84.5
Phi	GSM8K	High	.79	.02	.00	159.3
Phi	ARC	Low	.85	.00	.00	11.0
Phi	ARC	Med.	.87	.00	.00	11.0
Phi	ARC	High	.87	.00	.00	11.0
Phi	MMLU-Pro	Low	.20	.08	1.00	45.0
Phi	MMLU-Pro	Med.	.28	.06	1.00	49.8
Phi	MMLU-Pro	High	.31	.04	.00	114.3
Phi	MATH-500	Low	.40	.62	1.00	291.4
Phi	MATH-500	Med.	.40	.20	.74	291.4
Phi	MATH-500	High	.40	.09	.19	291.4

Table 7: Confirmatory strategy selections pooled across datasets (400 decisions per row).

Model	Tier	Direct	CoT	Few-Shot	Self-Refine	Critic-Refine
Qwen	Low	181	0	219	0	0
Qwen	Medium	256	0	144	0	0
Qwen	High	156	71	172	1	0
Phi	Low	100	0	300	0	0
Phi	Medium	157	0	243	0	0
Phi	High	217	1	141	41	0

Table 8: Offline full-information macro comparisons from the frozen exhaustive matrix. Values are accuracy/violation/mean tokens.

Model	Tier	Conservative V2	V1 point	Best fixed	Oracle
Qwen	Low	.603/.268/165.7	.590/.288/173.9	.675/.848/417.7	.833/.355/215.1
Qwen	Med.	.583/.125/195.7	.610/.218/244.9	.675/.503/417.7	.833/.140/215.1
Qwen	High	.593/.030/266.7	.648/.123/388.3	.675/.143/417.7	.833/.040/215.1
Phi	Low	.468/.225/99.1	.520/.210/83.3	.638/.738/443.0	.818/.383/151.5
Phi	Med.	.525/.068/108.2	.600/.140/157.2	.638/.538/443.0	.818/.160/151.5
Phi	High	.583/.038/144.4	.668/.120/289.6	.638/.250/443.0	.818/.035/151.5

Table 9: Qwen always-fixed baselines from the offline exhaustive D3 matrix. Each entry is accuracy/violation/mean output tokens; L/M/H use the frozen budget thresholds.

Strategy	Low	Medium	High
Direct	.600/.615/323.5	.600/.263/323.5	.600/.075/323.5
Chain-of-Thought	.648/.835/403.3	.648/.448/403.3	.648/.120/403.3
Few-Shot	.595/.330/178.0	.595/.133/178.0	.595/.030/178.0
Self-Refine	.675/.945/485.9	.675/.690/485.9	.675/.235/485.9
Critic-Refine	.570/1.000/919.1	.570/.975/919.1	.570/.843/919.1

Table 10: Phi always-fixed baselines from the offline exhaustive D3 matrix. Each entry is accuracy/violation/mean output tokens; L/M/H use the frozen budget thresholds.

Strategy	Low	Medium	High
Direct	.590/.380/152.3	.590/.123/152.3	.590/.043/152.3
Chain-of-Thought	.658/.985/374.9	.658/.660/374.9	.658/.170/374.9
Few-Shot	.418/.150/62.5	.418/.043/62.5	.418/.013/62.5
Self-Refine	.670/.988/381.5	.670/.665/381.5	.670/.170/381.5
Critic-Refine	.605/1.000/766.1	.605/.985/766.1	.605/.793/766.1

I Generation Safety-Cap Audit

Explicit per-call finish metadata was checked for all 8,842 immutable D1/D3 model calls. There were 133 `max_new_tokens` terminations (1.50%): 81 of 4,401 Qwen calls (1.84%) and 52 of 4,441 Phi calls (1.17%). By dataset, the counts were 12/2,200 for ARC-Challenge, 0/2,241 for GSM8K, 100/2,201 for MMLU-Pro, and 21/2,200 for MATH-500. Thus safety-cap censoring was concentrated in MMLU-Pro and is not assumed absent. The audit uses explicit `finish_reason`; token equality to the cap is only a consistency check.

Development-stage safety-cap sensitivity. Across Stage B/C, 101 of 6,400 development generation calls (1.58%) reached dataset-specific safety caps: 49 calls affected 46/2,400 estimator-fit strategy executions and 52 affected 45/1,600 risk-calibration executions. These observations remained in the frozen procedures used for point-model fitting and accuracy priors, and for conservative correction, estimator selection, and budget construction. A post-hoc rank audit proved all eight model–dataset p25/p50/p75 budget triplets invariant to arbitrary upward uncensoring. Three selected-estimator or numerical-margin cells—Qwen MMLU-Pro, Phi MMLU-Pro, and Phi MATH-500—were not provably invariant because the unobserved continuations cannot be recovered without imputation.

Using the same point-model form but excluding affected estimator-fit records changed selected-strategy identity in 58/2,400 tier-specific sealed-confirmatory routing decisions (2.42%) and fallback status in 32/2,400 (1.33%); excluding affected accuracy-prior observations alone changed neither. For low/medium/high tiers, macro accuracy differences were +0.00125/+0.00125/+0.00250 and violation differences were -0.00250/+0.00250/+0.00500. One marginal comparison changed direction: Qwen MATH-500 medium violation was 0.27 for V1, 0.26 for frozen V2, and 0.28 under point-model exclusion. No inference was rerun, no frozen artifact or confirmatory hypothesis was changed, and these analyses are post hoc and descriptive rather than confirmatory.

Table 11: Resource accounting over all sealed-confirmatory executions.

Model	D1	D3	Calls	Generated tok.	Generation h	Peak VRAM (GB)
Qwen	1200	2000	4401	1,171,857	24.77	7.38
Phi	1200	2000	4441	836,108	21.55	9.05
Combined	2400	4000	8842	2,007,965	46.32	9.05

Table 12: Frozen ablations I: macro accuracy/violation/fallback/mean tokens.

Model	Tier	Frozen V2	No complexity	No accuracy prior
Qwen	Low	.603/.268/.688/165.7	.603/.268/.688/165.1	.603/.268/.688/165.7
Qwen	Med.	.583/.125/.438/195.7	.603/.103/.195/165.1	.603/.105/.438/165.7
Qwen	High	.593/.030/.123/266.7	.668/.095/.000/355.0	.603/.025/.123/165.7
Phi	Low	.468/.225/.750/99.1	.468/.223/.750/98.5	.468/.225/.750/99.1
Phi	Med.	.525/.068/.435/108.2	.553/.068/.340/112.8	.468/.065/.435/99.1
Phi	High	.583/.038/.048/144.4	.625/.060/.008/186.8	.468/.028/.048/99.1

Table 13: Frozen ablations II: macro accuracy/violation/fallback/mean tokens.

Model	Tier	No correction	Pooled Q90	Strategy Q90
Qwen	Low	.590/.288/.190/173.9	.603/.268/.750/165.7	.610/.280/.688/169.3
Qwen	Med.	.610/.218/.070/244.9	.603/.105/.510/165.7	.588/.128/.410/200.0
Qwen	High	.648/.123/.003/388.3	.585/.028/.123/247.4	.603/.035/.068/269.4
Phi	Low	.520/.210/.408/83.3	.418/.150/1.000/62.5	.545/.265/.750/111.9
Phi	Med.	.600/.140/.045/157.2	.475/.045/.500/71.6	.555/.068/.185/113.8
Phi	High	.668/.120/.000/289.6	.535/.023/.188/108.8	.593/.048/.048/157.0

J Frozen Ablations

No-complexity removes the complexity coefficient and uses overall priors; no-prior ranks feasible candidates by predicted tokens and frozen order. No-conservative-correction reuses the point estimate. Pooled and strategy- conditioned Q90 variants retain every other rule. Tables 12–13 report every frozen variant without post-hoc tuning.

K Runtime and Resource Accounting

Qwen used 4,401 calls, 1,171,857 generated tokens, and 24.77 measured generation-hours; Phi used 4,441 calls, 836,108 tokens, and 21.55 hours. Peak allocated VRAM was 7.38 and 9.05 GB. These totals exclude model loading. Interrupted runs resumed by execution identity and are not scientific failures; exact uninterrupted wall time is not reconstructible and is not reported.

L Reproducibility Checklist

Every result is bound to immutable records, deterministic execution keys, frozen fingerprints, analysis scripts, and an evidence manifest. Derived analyses reproduce byte-for-byte from those records. This review submission provides the anonymous paper PDF only; no supplementary archive is attached.

M Additional AI Assistance Details

AI assistance covered literature discovery, methodology feedback, code editing, experiment orchestration support, manuscript drafting and polishing, and mathematical checking. Human authorization governed every protocol freeze, amendment, sealed-data transition, and final interpretation.