PRISM: PARETO ROUTING OVER INPUT STRUCTURE AND MODELS FOR BUDGETED TABLE QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Table question answering (Table QA) requires a model to derive an answer from a variably-structured table and a natural-language question. Although general-purpose LLMs, including GPT-class systems, and smaller language models (SLMs) can solve many such questions, but a fixed model and fixed serialization force a deployment to choose between accuracy, cost, and latency before the difficulty of each input is known. Thus, we introduce PRISM, *Pareto Routing over Input Structure and Models*, which treats the model–representation pair as the primary decision and uses a second, evidence-gated call only when the first response fails a prespecified acceptance screen. The screen combines answer availability, evidence consistency, and reported confidence; a deterministic quota bounds mean API use by 1.25 calls per question. We derive a risk-priced continuation boundary, conditional representation–model crossover guarantees, and a risk bound that accounts for denied recovery. Completed experiments reach 83.02% strict accuracy on HiTab and 78.04% denotation accuracy on WikiTableQuestions (WTQ). On 851 paired WTQ questions, PRISM yields 43.43% more correct answers per accounted dollar than Pro/CSV, while trading away 8.96 accuracy points and adding 2.99 seconds of logical latency. The results support selective recovery and make its resource trade-offs explicit; they do not claim calibrated uncertainty or universal Pareto dominance.

1 INTRODUCTION

Table question answering (Table QA) asks a system to answer a natural-language question from a table whose rows, columns, headers, and sometimes hierarchical metadata carry the evidence needed for the answer. The task appears in flat benchmarks such as WikiTableQuestions (WTQ) and hierarchical benchmarks such as HiTab (Pasupat & Liang, 2015; Cheng et al., 2022). In deployment, the difficulty is not only the reasoning operation: the same table can be serialized as CSV, Markdown, or HTML, and each view exposes structure differently to a decoder. A practical Table QA system therefore has to decide how much model capacity and how much structural information to spend on each question.

A substantial body of work already addresses this task. Specialized table models such as TaPas, TaBERT, and TAPEX encode table structure or executable operations, while newer systems study generalist table models and structure-aware reasoning (Herzig et al., 2020; Yin et al., 2020; Liu et al., 2022; Zhang et al., 2024; Mouravieff et al., 2025; Wang et al., 2024). General-purpose LLMs, including GPT-class services, can perform Table QA through prompting, and smaller or cheaper LMs provide attractive throughput and cost. Model-routing and cascade methods already exploit quality–cost differences between models (Chen et al., 2023b; Ding et al., 2024; Ong et al., 2025; Jitkrittum et al., 2025). Controlled representation studies further show that changing the serialization of the same table can change answers (Sui et al., 2024; Zhang et al., 2026b).

These methods leave four coupled challenges. First, a larger model can improve reasoning but increase input and output cost, while an SLM can be economical but fail on difficult arithmetic, lookup, or multi-step cases. Second, a representation that exposes hierarchy may help one decoder while inflating tokens or distracting another; therefore format choice and model choice cannot be

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.

optimized independently. Third, a first answer can be unavailable, weakly supported, or inconsistent even when a second attempt would recover the answer, so routing needs an acceptance rule rather than an unconditional retry. Fourth, every recovery call consumes a shared budget, and the online controller cannot use the reference answer to decide whether to continue. The central problem is thus a budgeted, evidence-conditioned allocation problem: choose an initial model–representation pair, decide whether its response is safe to accept, and spend a limited fallback opportunity where the expected recovery value justifies its cost and latency.

PRISM addresses this problem in four steps. (1) A local selector chooses a primary pair (m, r) from model and content-equivalent representation candidates using only permitted input features and frozen development scores. (2) The primary response is checked for answer availability, confidence, and grounded cell evidence; these checks form one acceptance screen rather than unrelated exception handlers. (3) Responses that fail the screen request continuation, but a deterministic prefix quota admits at most a fixed fraction of requests. (4) An admitted fallback runs once, and a fixed final-selection rule compares the primary and fallback evidence without consulting the gold answer. This design makes the research question explicit: when does an additional, structurally different computation improve Table QA enough to justify its resource use?

Our contributions are: (i) a content-controlled formulation of joint model–representation routing for Table QA with explicit cost, latency, and quota objectives; (ii) an evidence-conditioned cascade with a pathwise call bound and risk analysis that states its calibration and fallback assumptions; (iii) conditional crossover and Pareto criteria for deciding when a representation change can substitute for a larger model; and (iv) completed HiTab and WTQ evaluations, including a shared-primary retry control, that quantify selective recovery and its deployment-specific trade-offs. The empirical results establish where the measured continuation helps, while the matched controls and audit delimit what cannot be claimed about universal representation superiority or calibrated uncertainty.

2 Related Work

Routing and adaptive inference. FrugalGPT studies economical use of model cascades, while Hybrid LLM and RouteLLM learn quality–cost routing policies (Chen et al., 2023b; Ding et al., 2024; Ong et al., 2025). UniRoute considers transfer to previously unobserved models, and recent training-free routing work studies online serving without gradient-based router training (Jitkrittum et al., 2025; Wu & Silwal, 2025; Song et al., 2025). Cost-sensitive contextual decision making also connects to bandits with resource constraints (Li et al., 2010; Badanidiyuru et al., 2013). Broader routing already includes more than a model identifier: HAPS considers architecture and parameter choices, and R2-Router and LLMRouterBench extend reasoning-based routing and evaluation (Tian et al., 2026; Xue et al., 2026; Li et al., 2026). Cold-start profiling and history-aware routing further motivate declaring what information a selector receives (Xu et al., 2026; Zhang et al., 2026a). Our axis is specifically the serialization of the *same table content*, with accounting for input-length changes and sequential fallback calls.

Tables, structure, and representation. WikiTableQuestions and HiTab expose compositional reasoning over flat and hierarchical tables (Pasupat & Liang, 2015; Cheng et al., 2022). TAPAS, TaBERT, and TAPEX incorporate table structure or executable operations through pretraining (Herzig et al., 2020; Yin et al., 2020; Liu et al., 2022); TableLlama develops a generalist table model (Zhang et al., 2024). Table Meets LLM and controlled representation studies establish that input structure matters (Sui et al., 2024; Zhang et al., 2026b). Structural Deep Encoding studies structure-aware attention, whereas Chain-of-Table modifies tables along a reasoning trajectory (Mouravieff et al., 2025; Wang et al., 2024). Recent table reasoning research also investigates inference-time scaling and reinforcement learning (Yang et al., 2025b;a). PRISM does not claim the discovery of representation sensitivity; it studies its use as an inference allocation variable under content, call, and cost controls.

Speculation, verification, and selective prediction. Speculative decoding accelerates token generation with a draft model and a target-model verification rule (Leviathan et al., 2023; Chen et al., 2023a; Narasimhan et al., 2024; Wu et al., 2025). Our speculation is at the *answer* level and does not preserve a target model's output distribution. Self-reported uncertainty contains useful signals but lacks automated calibration (Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2024). Selec-

tive classification and conformal risk control provide a vocabulary for acceptance and risk (Chow, 1970; El-Yaniv & Wiener, 2010; Angelopoulos et al., 2025); self-correction studies motivate distinguishing externally checked evidence from unconstrained revision (Huang et al., 2024). Our verifier checks grounded cell values and limited arithmetic, not complete semantic correctness. The risk theorem consequently states the calibration and conditional fallback assumptions it needs.

3 Joint Decisions over Content-Equivalent Table Views

Let $X = (T, Q) \sim \mathcal{D}$ denote a table and question, with target answer $Y^\star$. A table is a cell grid together with available header ancestry and merge metadata. A representation $r \in \mathcal{R} = \{\mathrm{CSV}, \mathrm{MD}, \mathrm{HTML}\}$ maps it to a string $S_r(T)$. Models belong to $\mathcal{M} = \{\mathrm{Lite}, \mathrm{Flash}, \mathrm{Pro}\}$. An action $a = (m, r) \in \mathcal{A} = \mathcal{M} \times \mathcal{R}$ produces answer $\widehat{Y}_a$, confidence s_a, and evidence e_a.

Definition 1 (Content equivalence)**.** Representations are content-equivalent on a table domain if each has a deterministic decoder D_r such that $D_r(S_r(T)) = T$, including all cell strings and supplied structure. Synthetic delimiters and coordinates may differ, but no representation omits rows, adds answers, or receives extra annotations.

Our three serializers satisfy this criterion by round-trip verification. "Markdown text view" refers to a full-table serialization, not relevance-based row slicing; "HTML view" denotes markup supplied as text, not browser rendering or a vision model. WTQ provides no original hierarchy or merge graph, so its adapter adds only the known header row and coordinate structure. This distinguishes syntax effects from privileged information.

Proposition 1 (Information preservation does not imply decoder invariance)**.** *For content-equivalent deterministic S_r, whenever the conditional mutual information is defined,*

$$I(Y^\star; S_r(T) \mid Q) = I(Y^\star; T \mid Q). \tag{1}$$

Consequently, differences in a fixed model's accuracy cannot be explained by an increase in the table's information about the answer alone. (Proof in Appendix A.)

For a fixed serving regime, define $p_a(x) = \mathbb{P}(\widehat{Y}_a = Y^\star \mid X = x)$, expected billed cost $c_a(x)$, and expected service latency $\ell_a(x)$. Equality denotes the chosen benchmark scoring rule. Randomness includes API behavior and generation, even when temperature is zero. Serving conditions are part of the estimand: changing concurrent load can change ℓ_a.

Definition 2 (Pareto routing objective)**.** A policy π maps available information to an initial action and a possible single escalation. We evaluate the deterministic policy class induced by the frozen selector, gate, request order, and final-selection rule; randomized mixtures of separately run policies are outside this empirical policy class. Its operating point is (A_π, C_π, L_π), with accuracy maximized and cost and latency minimized. Policy π dominates π' if $A_\pi \geq A_{\pi'}$, $C_\pi \leq C_{\pi'}$, and $L_\pi \leq L_{\pi'}$, with at least one strict inequality. The constrained objective is

$$\max_{\pi \in \Pi} A_\pi \quad \text{subject to} \quad C_\pi \leq B_C, \quad L_\pi \leq B_L, \quad \sum_{i=1}^{N} G_i \leq \lfloor \rho N \rfloor, \tag{2}$$

where G_i indicates an admitted escalation and $\rho < 1$ is fixed before execution.

The scalar utility $u_a(x) = p_a(x) - \lambda c_a(x) - \mu \ell_a(x)$, $\lambda, \mu \geq 0$, is a useful local decision surrogate. It does not assert that a weighted sum recovers every nondominated point of a discrete, nonconvex action set (Ehrgott, 2005). We report observed operating points in addition to aggregate frontier measures; the reported Pareto statements refer to this prespecified deterministic policy family.

Definition 3 (Evidence-conditioned decision risk)**.** Let $\mathcal{I} = (X, O, \mathcal{F})$ contain the input, primary observation and permitted predecision history. Define $\mathcal{L}_0 = 1 - Y_0$ for returning the primary and $\mathcal{L}_1 = 1 - Y_1$ for continuation followed by the fixed final-selection rule. A missing answer incurs loss one under the benchmark protocol. The conditional decision risks are $r_j(\mathcal{I}) = \mathbb{E}[\mathcal{L}_j \mid \mathcal{I}]$. A conservative envelope $\overline{r}_0(\mathcal{I}) = \sup_{P \in \mathcal{Q}(\mathcal{I})} \mathbb{E}_P[\mathcal{L}_0 \mid \mathcal{I}]$ certifies acceptance at tolerance γ only if the true conditional law belongs to the ambiguity set $\mathcal{Q}(\mathcal{I})$ and $\overline{r}_0 \leq \gamma$.

Proposition 2 (Risk-priced continuation boundary)**.** *With known conditional risks and expected incremental resources, the relaxed binary decision purchases continuation exactly when*

$$r_0(\mathcal{I}) - r_1(\mathcal{I}) > \lambda c_1(\mathcal{I}) + \mu \ell_1^{\mathrm{time}}(\mathcal{I}) + \eta, \tag{3}$$

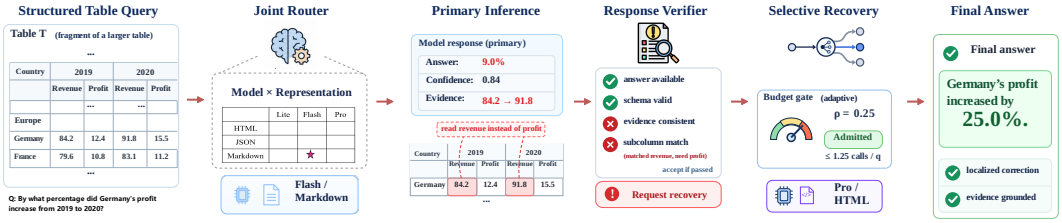


Figure 1: **Overview of the PRISM framework.** Given a structured table query, the joint router selects an initial model–representation pair (e.g., Flash/Markdown). The response verifier checks the primary response for answer availability, schema validity, and grounded cell evidence. When an anomaly is detected, an evidence-gated recovery request is admitted under an adaptive budget quota ($\rho = 0.25$, bounded by ≤ 1.25 calls/q) and escalated to a high-capacity configuration (Pro/HTML) for grounded recovery.

where $\eta \geq 0$ prices one continuation slot; ties may be randomized. (Proof in Appendix A.)

This boundary prices the value of recovery, not merely the presence of uncertainty. A failed acceptance certificate motivates a request; it does not prove that continuation is beneficial or that a constrained budget can admit it.

4 PRISM: Joint Selection and Budgeted Verification

4.1 Evidence-conditioned risk minimization

Problem & Challenge. After a primary Table QA call, the controller must decide whether the returned answer is sufficiently supported to accept or whether another call is worth purchasing. Self-reported confidence alone cannot resolve this decision because an answer can be missing, malformed, or inconsistent with cited cells. Moreover, the gate sees only the primary response and the remaining quota without access to reference answers or calibrated uncertainty.

Design. As shown in Figure 1, PRISM decomposes execution into primary configuration, one unified risk screen, and resource admission. An unavailable answer, an incomplete parse, inconsistent evidence, and missing confidence all fail the same certificate. Let $V(O) = 1$ when an answer is available, confidence s is finite and the evidence check is consistent. The deployed score is

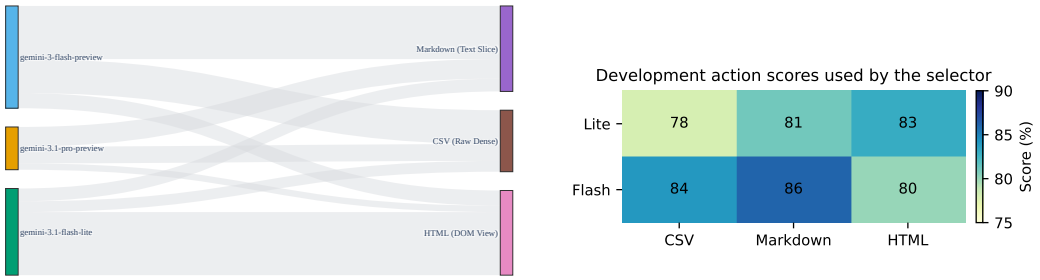
$$\widehat{r}_0(O) = \begin{cases} 1 - s, & V(O) = 1, \\ 1, & V(O) = 0, \end{cases} \qquad Z = \mathbf{1}\{\widehat{r}_0 > \gamma\}, \quad \gamma = 1 - \tau = .2. \tag{4}$$

Equation (4) exactly reproduces the frozen gate. An anomaly deterministically crosses the *screening* threshold. For unavailable output the return is scored wrong; for other failed checks, one is a conservative uncertainty score, not a measured probability of semantic error. In ambiguity-set terms, lack of a certificate permits worst-case loss one. Treating $1 - s$ as an upper risk bound on the accepted branch additionally requires conditional calibration; the current experiments do not establish that premise. Consequently, these signals express insufficient evidence to accept, without identifying a Bayesian decomposition into epistemic and aleatoric uncertainty.

When requested recovery is admitted, Pro/HTML supplies a new WTQ candidate (Pro/CSV in the locked HiTab run). This is a feasible surrogate for Proposition 2: schema/evidence checks approximate the sign of recovery value, and the prefix quota enforces resources independently of that approximation. Writing $d(O) = \mathbb{E}[Y_1 - Y_0 \mid O]$, its ideal relaxed rule is

$$d(O) - \lambda c_1(O) - \mu \ell_1(O) > \eta. \tag{5}$$

The exact opportunity cost of a slot depends on horizon and history; constant η is a relaxation. The deployed controller uses local scores and no paid assessor or gradient update. It does not claim to estimate conditional recovery value or solve the optimal controller. The implemented selector ranks primary actions myopically; Appendix A.11 bounds the loss from omitting continuation value.



(a) Observed primary allocation

(b) Frozen development action scores

Figure 2: **Primary model–representation selection in the frozen WTQ protocol. (a)** Observed primary allocation from model choices to content-equivalent table views; arrows denote policy support. **(b)** Frozen development action scores used by the WTQ selector (prospective ranking inputs).

4.2 Local selection and representation crossover

Problem & Challenge. The initial call must choose both a model and a serialization before any response evidence is available. A single-axis selector cannot trade structural exposure against model capacity, while the heterogeneous action space demands an efficient, label-free decision rule that avoids overfitting WTQ answers.

Design. The initial selector uses table size and question-only intent to index a small context partition. The deployed WTQ primary candidate set comprises six Lite/Flash–format actions; Pro is reserved for continuation. The locked HiTab primary set contains only Lite/CSV and Flash/CSV. These are explicit restrictions of the nine-action formulation. In the development-calibrated instantiation, empirical action counts are shrunk toward a global success estimate:

$$\widetilde{p}_a(x) = \frac{k_{b(x),a} + \kappa(k_a+1)/(n_a+2)}{n_{b(x),a} + \kappa}, \qquad \widehat{a}(x) \in \arg\max_a \{\widetilde{p}_a(x) + b_a(x) - \lambda \widehat{c}_a(x)\}. \tag{6}$$

Here $b(x)$ is the context bucket, $\kappa = 30$, and $b_a(x)$ is an explicitly versioned structural preference. With $b_a = 0$, the score is the posterior mean of a binomial likelihood under a plug-in Beta prior with parameters $\kappa\bar{p}_a$ and $\kappa(1-\bar{p}_a)$, where $\bar{p}_a = (k_a+1)/(n_a+2)$. This empirical-Bayes-style shrinkage does not propagate uncertainty in the global estimate. Once the hand-set preference $b_a(x)$ or a cost term is added, the expression is a heuristic ranking score, not a posterior probability or a pure posterior-expected-utility calculation; it may lie outside $[0,1]$. HiTab uses development-derived counts; the WTQ policy uses fixed synthetic priors and preferences without fitting WTQ answer labels. Both avoid gradients, but only the latter is task-label-free with respect to WTQ.

The following result formalizes a substitution opportunity without assuming that representations are universally better than larger models. At a context x, take a reference action $a_0 = (m, r)$, a representation alternative $a_R = (m, r')$, and a model alternative $a_M = (m', r)$. Define $\Delta p_j = p_{a_j} - p_{a_0}$ and analogously $\Delta c_j, \Delta \ell_j$.

Theorem 1 (Crossover and representation dominance certificate)**.** *Suppose valid bounds satisfy* $\Delta p_j \in [\underline{g}_j, \overline{g}_j]$, $\Delta c_j \in [\underline{d}_j, \overline{d}_j]$, *and* $\Delta \ell_j \in [\underline{t}_j, \overline{t}_j]$ *for* $j \in \{R, M\}$. *If*

$$\underline{g}_R - \overline{g}_M > \lambda(\overline{d}_R - \underline{d}_M) + \mu(\overline{t}_R - \underline{t}_M), \tag{7}$$

then $u_{a_R} > u_{a_M}$. *If instead* $\underline{g}_R \geq \overline{g}_M$, $\overline{d}_R \leq \underline{d}_M$, *and* $\overline{t}_R \leq \underline{t}_M$, *with a strict inequality, then* a_R *Pareto-dominates* a_M *at* x. *If* $0 < \underline{d}_j \leq \Delta c_j \leq \overline{d}_j$ *and* $0 \leq \underline{g}_j \leq \Delta p_j \leq \overline{g}_j$ *for both alternatives,*

$$\frac{\underline{g}_R}{\overline{d}_R} > \frac{\overline{g}_M}{\underline{d}_M} \quad \Longrightarrow \quad \frac{\Delta p_R}{\Delta c_R} > \frac{\Delta p_M}{\Delta c_M}. \tag{8}$$

(Proof in Appendix A.)

The crossover separates local conversion time from model-side token inflation. With configured input/output prices $q_m^{\text{in}}, q_m^{\text{out}}$ per million tokens, the alternative-action cost difference is

$$\Delta c_R = 10^{-6}\left(q_m^{\text{in}} \Delta n_R^{\text{in}} + q_m^{\text{out}} \Delta n_R^{\text{out}}\right). \tag{9}$$

Local work contributes to latency; it is not an API token charge. Both output-token change and HTML's additional markup belong in this comparison. An already-started first call is a sunk cost: an escalation is charged in full. Expanding the action set guarantees non-negative utility headroom under bounded estimation error and selection overhead (Theorem 3 in Appendix A), formalizing why expanding the candidate set can pay off provided evaluation overhead does not erase the margin.

Figure 2(b) visualizes these frozen development scores across model–representation pairs, while Figure 2(a) displays the realized primary allocation after context preferences and tie-breaking. Feasible stage-specific supports restrict the full nine-action catalog: locked HiTab ranks two CSV primaries, while WTQ ranks six cheap-model configurations and reserves Pro/HTML for continuation (Figure 2(a); Appendix A.12).

4.3 Evidence-gated escalation with a pathwise budget

Problem & Challenge. A failed certificate should trigger continuation only when the shared budget can pay for it. Requests arrive in an online stream where calls can fail independently, requiring the controller to enforce a strict quota with at most one fallback attempt per question.

Design. We structure staged execution to commit an initial computation, inspect its returned evidence, and purchase a secondary fallback configuration only when admitted by the verification gate under standard per-call token limits. The primary response supplies its answer, $s_i \in [0, 1]$, and up to 16 cell citations. The verifier checks coordinates, exact cited cell text, and supported lookup or arithmetic consistency. It cannot verify that cited operands answer the right question or exhaust the relevant rows. Invalid or incomplete output, missing/low confidence, and unsupported or inconsistent evidence all request escalation, denoted $Z_i = 1$.

Let $K_{i-1} = \sum_{j<i} G_j$. Requests are processed in a fixed shuffled order and admitted by

$$G_i = Z_i \mathbf{1}\{K_{i-1} < \lfloor \rho i \rfloor\}, \qquad \rho = 0.25. \tag{10}$$

When admitted, a single Pro call produces a replacement candidate. Transport or schema failure preserves a usable primary answer; a known-inconsistent replacement does not override a primary answer with consistent evidence. This final-selection rule is part of the policy, not an oracle choice between two answers.

Theorem 2 (Deterministic call bound)**.** *If each primary and admitted fallback starts at most one HTTP attempt, then every prefix satisfies $\sum_{j \le i} G_j \le \lfloor \rho i \rfloor$ and the full run starts at most $N + \lfloor \rho N \rfloor$ attempts. Each question starts at most two attempts, regardless of failures or dependence among responses. (Proof in Appendix A.)*

4.4 Recovery accounting and conditional risk control

Problem & Challenge. The value of continuation is a selective treatment effect depending on which responses fail screening and whether fallback corrects or harms them. Since accepted, denied, and granted cases follow different mechanisms, risk accounting must explicitly reflect this partition.

Design. Write Y_0 for primary correctness and Y_1 for correctness after the fallback and final-selection rule, on escalated examples. The final correctness is $Y_\pi = (1-G)Y_0 + GY_1$. Without independence assumptions,

$$A_\pi - A_0 = \mathbb{E}[G(Y_1 - Y_0)] = \mathbb{P}(G, Y_0 = 0, Y_1 = 1) - \mathbb{P}(G, Y_0 = 1, Y_1 = 0). \tag{11}$$

Thus selective rescues, rather than the size of the fallback model, determine net benefit. Cost is $C_\pi = C_0 + \mathbb{E}[GC_1]$, where C_1 is the entire extra call. Partitioning executions into accepted primaries $S = \{Z = 0\}$, denied requests $D = \{Z = 1, G = 0\}$, and granted recovery $G = 1$ yields an idealized risk bound $1 - A_\pi \le \gamma \mathbb{P}(S) + \mathbb{P}(D) + r_f \mathbb{P}(G)$ (Theorem 4, Appendix A). Because self-reported confidence is uncalibrated, this remains an idealized benchmark rather than an empirical certificate, making the risk penalty of quota denials explicit. Appendix A gives the complete proof, an ideal resource-priced gate, and counterexamples to unconditional dominance.

5 Experimental Protocol

Completed benchmarks and frozen policies. HiTab contains 1,584 corrected test questions from 538 tables. A 150-question development pilot calibrated local scores and selected CSV-only primary

Table 1: Completed full benchmarks. HiTab uses strict accuracy; WTQ uses official denotation accuracy. Costs include unknown-billing reserves. Times are the original wave-based service measurements.

Benchmark	Policy	Accuracy (%)	USD/q	Mean s	Calls/q
HiTab (1,584)	Primary	78.47	.003152	5.78	1.000
	Joint PRISM	83.02	.006462	14.56	1.237
WTQ (4,344)	Primary	70.72	.003283	10.44	1.000
	Joint PRISM	78.04	.007373	30.02	1.250

support. Eight tables containing 33 questions were exposed in an earlier diagnostic; the unexposed cohort has 1,551 questions. WTQ contains 4,344 unseen-table questions from 421 tables. Its frozen priors select Lite/HTML, Flash/Markdown, or Flash/CSV; no WTQ answer labels fit these priors. HiTab escalates to Pro/CSV and WTQ to Pro/HTML. Both use 2,048-token ceilings, threshold $\tau = .8$, quota $\rho = .25$, and at most one attempt per stage. We distinguish gradient-free HiTab calibration from task-label-free WTQ priors.

Frozen follow-up and controls. After inspecting the full runs, we froze a stratified follow-up before any new inference: 851 WTQ questions from 180 tables for paired PRISM and Pro/CSV, with nested cohorts for restricted selectors and format swaps. This is an adaptive follow-up, not an untouched replication. The model-only selector fixes CSV and considers the two cheap models; its frozen scores select Flash on all 572 questions. The representation-only selector fixes Flash for its primary; both restrictions retain Pro/HTML continuation. They therefore test primary selection restrictions, not entirely single-model serving. A 380-question swap keeps the model and changes HTML or Markdown to CSV, and CSV to Markdown. Every one of 212 new PRISM grants also receives an independent same-model/same-format shadow continuation. The deployed and shadow policies share the primary, gate and final-selection rule; each is charged only its own continuation. HiTab contributes 140 same-model CSV/HTML pairs; 333 historical WTQ rescues and 85 Pro/Markdown questions are separately labeled diagnostics. Appendix C gives the full allocation.

Endpoints, cost, and latency. HiTab's primary endpoint is normalized ordered-answer strict correctness. WTQ uses the official v1.0.2 denotation evaluator with canonical targets; invalid answers and transport failures remain in the denominator. Cost is USD per question over all policy calls using configured rates (Table 1); unknown bills receive disclosed reserves. Service latency in original runs measures wave-based time, while the follow-up measures logical request time (excluding synchronization; see Appendix B).

Paired inference. The follow-up samples tables within cell-count strata (up to 5 questions per WTQ table, 2 for HiTab). We report inverse-inclusion-weighted ratio estimates using narrower-cohort weights on both contrast sides. Intervals use 2,000 paired stratified table-cluster resamples, with inverse-inclusion weighting as described in Appendix C; three primary contrasts add Bonferroni adjustments. Appendix C details weights and frontier summaries. Full-run gains use paired table resampling without survey weights.

6 Results: Recovery, Representation, and Resource Allocation

6.1 Full runs establish bounded recovery

HiTab improves from 1,243 to 1,315 correct answers: 83 rescues minus 11 harms yield 72 net corrections, or 4.55 points (95% CI $[3.11, 5.99]$). On unexposed tables, accuracy is 82.98%, with gain 4.65 points $[3.20, 6.13]$. WTQ improves from 3,072 to 3,390 correct answers: 333 rescues minus 15 harms yield 7.32 points $[6.47, 8.21]$. The gate grants 376/420 requested HiTab escalations and 1,085/2,356 WTQ requests. Starts total 1,960 and 5,429, below the respective ceilings 1,980 and 5,430. These paired gains establish the value of the observed continuation relative to the same first responses; they do not compare against an independently optimized primary.

Recovery from output/schema or transport problems is a substantial part of the observed gain, but the updated official-rescore totals are 83 rescues and 11 harms on HiTab and 333 rescues and 15

Table 2: Where recovery has impact. Cohorts overlap and were analyzed retrospectively using existing structure and gate criteria, without selecting on correctness. WTQ uses denotation and HiTab strict accuracy; only fresh WTQ rows are survey-weighted. Intervals are descriptive pointwise 95%, not multiplicity-adjusted subgroup discoveries. R/H gives raw rescues/harms.

Cohort	N	Primary %	Final %	Gain pp [95% CI]	R/H
WTQ: requested	2356	53.61	67.11	13.50 [12.02, 15.00]	333/15
WTQ: admitted	1085	53.92	83.23	29.31 [26.31, 32.36]	333/15
Fresh WTQ: admitted	212	49.77	81.44	31.67 [25.18, 38.34]	73/7
WTQ: >300 cells	347	63.69	69.16	5.48 [2.74, 8.17]	20/1
Fresh WTQ: >300 cells	80	63.21	72.37	9.16 [2.81, 16.57]	8/2
HiTab: depth ≥ 3	838	76.01	80.67	4.65 [2.88, 6.67]	43/4

harms on WTQ. The remaining corrections include calculation mismatches and table-grounding or semantic-evidence errors, showing that the continuation is not limited to syntax repair. In particular, on WTQ, correcting numerical calculation errors alone accounts for 85 net gains. An additional 37 parseable incomplete responses also reach the limit. All 769 length-terminated primaries are Flash calls, with a different question mix from Lite. Self-confidence is concentrated near one: only seven finite WTQ scores are below .8, and none in HiTab. The operational gate is primarily a schema/evidence detector, not an empirically calibrated confidence test. Appendix D reports exact confusion counts and failure reasons.

6.2 Pareto efficiency and cost-normalized utility

PRISM gains 7.74 points $[6.03, 9.58]$ over its fresh primary while using 37.64% less accounted cost than Pro/CSV. It returns 93.97 expected correct answers per dollar versus Pro's 65.52, a 43.43% relative increase (exploratory paired interval $[32.26, 56.09]\%$). This productivity measure rewards resource efficiency; it does not imply equal fidelity. The accuracy difference is -8.96 points, with paired interval $[-11.68, -6.21]$ and family-adjusted interval $[-12.28, -5.57]$. PRISM's logical latency is 2.99 seconds higher $[1.65, 4.38]$. Pro therefore anchors high observed accuracy at a higher price, but not a higher measured latency.

These points represent distinct deployment trade-offs. For $U = A - \lambda C - \mu L$, their empirical pairwise utility difference is $\widehat{U}_{\mathrm{PRISM}} - \widehat{U}_{\mathrm{Pro}} = -.089616 + .004877\lambda - 2.99243\mu$. At $\mu = 0$, PRISM is preferable to Pro alone when $\lambda > 18.38$; other observed policies can offer greater utility. At the prespecified \$0.05 reference ceiling, policy-alone cost AUC is .6366 for PRISM and .6289 for Pro, a difference .0077 with interval $[-.0154, .0301]$. This is an uncertain point advantage, not decisive AUC dominance. All four master-cohort points are Pareto-nondominated. Appendix F reports the complete utility family and avoids conflating per-dollar productivity, AUC and Pareto dominance.

6.3 High-impact recovery and structural cohorts

The recovery boundary concentrates added computation where primary evidence is inadequate (Table 2). On the original WTQ run, the 1,085 admitted cases improve from 53.92% to 83.23% accuracy: $+29.31$ points $[26.31, 32.36]$, or 54.36% relative improvement, with a rescue-to-harm ratio of $333 : 15 = 22.2 : 1$. This is the ratio for the *admitted-recovery cohort*, not a ratio inferred for deep nesting or for every structurally difficult input. Keeping denied cases in the requested cohort gives a smaller, deployment-relevant gain of 13.50 points on 2,356 questions. Admission therefore determines how much potential recovery can be realized under the fixed budget.

The fresh run reproduces concentrated recovery: admitted-case weighted accuracy rises from 49.77% to 81.44%, with 73 rescues and seven harms. Crucially, the same 212 admissions also have a same-model shadow response. PRISM exceeds that matched alternative by 15.16 points $[9.35, 21.35]$. Because both alternatives share the primary response and admission decision, this comparison attributes additional recovery to the Pro/HTML continuation package, beyond an extra attempt alone. It does not isolate representation from model capability differences.

Structural complexity: favorable estimates, limited resolution. We use existing metadata to define large tables (> 300 cells) and deeply nested HiTab tables (header depth ≥ 3; WTQ provides flat headers only). HiTab's 838 deep-table questions gain 4.65 points $[2.88, 6.67]$ with 43 rescues and four harms. Fresh large-table WTQ gains 9.16 points $[2.81, 16.57]$ over its primary; PRISM also exceeds shared-primary retry by 11.15 points $[4.65, 18.51]$ on these 80 questions. Against model-only and format-only selectors, large-table paired advantages are 8.41 and 12.81 points (53 questions from 11 tables per comparison), though intervals include zero. Furthermore, borderline confidence ($0.8 \leq s < .95$) gains 9.09 points on original WTQ but loses 9.61 weighted points on fresh samples, confirming that self-reported scores are unstable across shifts and that recovery primarily rescues unavailable or syntactically invalid outputs (Appendix F).

6.4 Continuation Value over Homogeneous Retry

The matched same-model continuation reaches 72.23% accuracy, versus PRISM's 75.93% (paired difference $+3.70$ points $[2.22, 5.22]$). Relative to the shared primary, PRISM produces 73 rescues and 7 harms, while same-model retry produces 37 rescues and 2 harms. PRISM adds substantial recovery beyond cheap repetition at an extra \$0.003153/q, while achieving 1.27s lower logical latency (15.56s vs 16.83s; Table 6) because Pro over HTML resolves coordinates in fewer decoding steps. Because both arms share identical primaries and gate decisions, the comparison confirms that same-model repetition frequently repeats identical structural misreadings, whereas heterogeneous escalation over both syntax and model tiers delivers superior recovery and throughput without the instability of unguided multi-turn sampling.

7 Discussion

PRISM connects joint model–representation choice to evidence-conditioned acceptance and bounded recovery. Its strongest demonstrated capability is concentrated continuation: large accuracy gains on admitted uncertain responses, an empirical $22.2 : 1$ rescue-to-harm ratio on WTQ, and clear superiority over homogeneous retry. Structural cohorts further highlight where recovery provides the highest payoff, while cost-normalized productivity demonstrates resource efficiency without obscuring underlying accuracy and latency trade-offs.

Deployment trade-offs. For budget-constrained applications ($B \leq \$0.01$/q), PRISM routes standard queries to lightweight SLMs while reserving frontier capacity for verifiable reasoning failures. Conversely, when strict accuracy is paramount and budgets permit ($B \geq \$0.05$/q), unconstrained Pro/CSV remains superior, as PRISM trades away 8.96 accuracy points on WTQ to achieve a 37.64% cost reduction. For intermediate regimes, practitioners can dynamically adjust the prefix quota ρ to trace a tunable accuracy–cost frontier under strict latency SLAs.

Delimitations. Pro/CSV achieves higher absolute fidelity, and format swaps show that representation preferences are context-dependent rather than globally dominant. Retrospective structural estimates remain noisy, and self-reported model confidence is not an epistemic calibration guarantee. Hand-designed selectors, reliance on a single model family, potential API backend drift, and benchmark provenance further constrain generalization. Operating under fixed quotas and finite action spaces, PRISM delimits where selective recovery merits additional computation rather than asserting universal Pareto dominance or uniform risk bounds.

8 Conclusion

We presented PRISM, a budgeted framework for Table QA that jointly optimizes model capacity and table serialization before execution, and selectively escalates uncertain responses under a deterministic quota. By combining content-equivalent representations, grounded verifiers, and pathwise call bounds, PRISM achieves 83.02% strict accuracy on HiTab and 78.04% on WTQ within a deterministic ≤ 1.25 calls-per-question ceiling, delivering 43.43% more correct answers per dollar than static frontier calls, all while avoiding the resource overhead of unconstrained test-time rollouts. The resulting method converts structural heterogeneity into measurable efficiency gains, offering an auditable, cost-effective paradigm for reliable tabular reasoning under strict resource constraints.

AI USE STATEMENT

Generative AI tools assisted with research planning, software implementation, experimental analysis, literature organization, mathematical exposition, and manuscript drafting. The evaluated language models generated benchmark answers. Automated validation checked the final archive hashes, reproduced the frozen analyses, tested serializer round trips and evaluator behavior, and checked manuscript source consistency. This describes the automated work performed; it does not certify an independent human audit. Responsibility for submission, scientific claims, and the final scope of review rests with the submitting authors.

ETHICS STATEMENT

This work uses existing TableQA benchmarks and commercial model APIs, without recruiting human participants or making decisions about individuals. We report negative comparisons, failed outputs, sample-selection procedures, and unknown billing. Public benchmark provenance does not establish freedom from model pretraining contamination. Dataset redistribution and model usage remain subject to their respective terms. The system's confidence and evidence checks are not guarantees of semantic correctness or suitability for consequential financial decisions.

REPRODUCIBILITY STATEMENT

Appendices A–G give proofs, frozen parameters, sampling weights, endpoints, and audit results. The accompanying artifacts contain frozen code, tasks, predictions, attempt ledgers, canonical targets, paired statistics, and hashes. Offline reproduction uses saved outputs and requires no additional model queries. Original full-run and final follow-up results remain separate, with each latency definition and scoring reconciliation disclosed. The final analysis reproduces all archived summary tables byte-for-byte; additional frontier uncertainty is labeled exploratory.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control, 2025. URL https://arxiv.org/abs/2208.02814.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 207–216. IEEE, October 2013. doi: 10.1109/focs.2013.30. URL http://dx.doi.org/10.1109/FOCS.2013.30.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling, 2023a. URL https://arxiv.org/abs/2302.01318.

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

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. Hitab: A hierarchical table dataset for question answering and natural language generation, 2022. URL https://arxiv.org/abs/2108.06712.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of information theory*, volume 2. wiley New York, 1991.

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

Miroslav Dudik, John Langford, and Lihong Li. Doubly robust policy evaluation and learning, 2011. URL https://arxiv.org/abs/1103.4601.

Matthias Ehrgott. *Multicriteria optimization*. Springer, 2005.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL http://jmlr.org/papers/v11/el-yaniv10a.html.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. TaPas: Weakly supervised table parsing via pre-training. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4320–4333, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.398. URL https://aclanthology.org/2020.acl-main.398/.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet, 2024. URL https://arxiv.org/abs/2310.01798.

Wittawat Jitkrittum, Harikrishna Narasimhan, Ankit Singh Rawat, Jeevesh Juneja, Congchao Wang, Zifeng Wang, Alec Go, Chen-Yu Lee, Pradeep Shenoy, Rina Panigrahy, Aditya Krishna Menon, and Sanjiv Kumar. Universal model routing for efficient llm inference, 2025. URL https://arxiv.org/abs/2502.08773.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding, 2023. URL https://arxiv.org/abs/2211.17192.

Hao Li, Yiqun Zhang, Zhaoyan Guo, Chenxu Wang, Shengji Tang, Qiaosheng Zhang, Yang Chen, Biqing Qi, Peng Ye, Lei Bai, Zhen Wang, and Shuyue Hu. Llmrouterbench: A massive benchmark and unified framework for llm routing, 2026. URL https://arxiv.org/abs/2601.07206.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pp. 661–670. ACM, April 2010. doi: 10.1145/1772690.1772758. URL http://dx.doi.org/10.1145/1772690.1772758.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. Tapex: Table pre-training via learning a neural sql executor, 2022. URL https://arxiv.org/abs/2107.07653.

Raphaël Mouravieff, Benjamin Piwowarski, and Sylvain Lamprier. Structural deep encoding for table question answering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2389–2402, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.121. URL https://aclanthology.org/2025.findings-acl.121/.

Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Seungyeon Kim, Neha Gupta, Aditya Krishna Menon, and Sanjiv Kumar. Faster cascades via speculative decoding, 2024. URL https://arxiv.org/abs/2405.19261.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data, 2025. URL https://arxiv.org/abs/2406.18665.

Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In Chengqing Zong and Michael Strube (eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1470–1480, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1142. URL https://aclanthology.org/P15-1142/.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

Wei Song, Zhenya Huang, Cheng Cheng, Weibo Gao, Bihan Xu, GuanHao Zhao, Fei Wang, and Runze Wu. Irt-router: Effective and interpretable multi-llm routing via item response theory, 2025. URL https://arxiv.org/abs/2506.01048.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study, 2024. URL https://arxiv.org/abs/2305.13062.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, 2023. URL https://arxiv.org/abs/2305.14975.

Zihang Tian, Rui Li, Jingsen Zhang, Xiaohe Bo, Wei Huo, and Xu Chen. Haps: Hierarchical llm routing with joint architecture and parameter search, 2026. URL https://arxiv.org/abs/2601.05903.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. Chain-of-table: Evolving tables in the reasoning chain for table understanding, 2024. URL https://arxiv.org/abs/2401.04398.

Fangzhou Wu and Sandeep Silwal. Efficient training-free online routing for high-volume multi-llm serving, 2025. URL https://arxiv.org/abs/2509.02718.

Hang Wu, Jianian Zhu, Yinghui Li, Haojie Wang, Biao Hou, and Jidong Zhai. Specrouter: Adaptive routing for multi-level speculative decoding in large language models, 2025. URL https://arxiv.org/abs/2505.07680.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms, 2024. URL https://arxiv.org/abs/2306.13063.

Jingjun Xu, Hongji Pu, Tao Feng, Haozhen Zhang, Jiaxuan You, and Ge Liu. Routeprofile: Graph-based profiling for cold-start llm routing, 2026. URL https://arxiv.org/abs/2605.00180.

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

Saisai Yang, Qingyi Huang, Jing Yuan, Liangyu Zha, Kai Tang, Yuhang Yang, Ning Wang, Yucheng Wei, Liyao Li, Wentao Ye, Hao Chen, Tao Zhang, Junlin Zhou, Haobo Wang, Gang Chen, and Junbo Zhao. Tablegpt-r1: Advancing tabular reasoning through reinforcement learning, 2025a. URL https://arxiv.org/abs/2512.20312.

Zheyuan Yang, Lyuhao Chen, Arman Cohan, and Yilun Zhao. Table-r1: Inference-time scaling for table reasoning, 2025b. URL https://arxiv.org/abs/2505.23621.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. TaBERT: Pretraining for joint understanding of textual and tabular data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8413–8426, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.745. URL https://aclanthology.org/2020.acl-main.745/.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. Tablellama: Towards open large generalist models for tables, 2024. URL https://arxiv.org/abs/2311.09206.

Yiqun Zhang, Hao Li, Zihan Wang, Shi Feng, Xiaocui Yang, Daling Wang, Bo Zhang, Lei Bai, and Shuyue Hu. Mtrouter: Cost-aware multi-turn llm routing with history-model joint embeddings, 2026a. URL https://arxiv.org/abs/2604.23530.

Yue Zhang, Seiji Maekawa, and Nikita Bhutani. Same content, different representations: A controlled study for table qa, 2026b. URL https://arxiv.org/abs/2509.22983.

A Complete Derivations and Assumption Boundaries

A.1 Notation and policy information

For each context x, the reference action a_0, representation alternative a_R, and model alternative a_M are alternative *first* executions. Their potential outcomes need not be observed together. Let $\mathcal{F}_{i-1}$ contain previous requests, gate grants, billing, and allowed observable responses, but no reference answers used for online adaptation. The primary selector observes X_i and the frozen policy. The gate additionally observes the first response's schema status, confidence, evidence and computation checks. A fallback sees the table/question under its own serialization; the current implementation does not append the primary answer to its prompt. Consequently, it is a fresh attempt rather than an explicitly conditioned correction prompt.

Table 3: Principal quantities and their interpretations.

Symbol	Meaning
S_r, D_r	Serializer and left inverse on the supported table domain
p_a, c_a, ℓ_a	Expected correctness, billed cost, and service latency for action a
u_a	Scalar utility $p_a - \lambda c_a - \mu \ell_a$
Z_i, G_i	Requested and admitted escalation indicators
S, D	Accepted-primary and requested-but-denied events
Y_0, Y_1, Y_π	Primary, post-escalation final, and returned correctness
ρ, τ	Escalation fraction cap and self-confidence threshold
ϵ, h	Uniform utility approximation error and selection overhead

A.2 Proof of Proposition 1

Condition on Q. Since $S_r(T)$ is a deterministic function of T, data processing gives $I(Y^\star; S_r(T) \mid Q) \le I(Y^\star; T \mid Q)$. Since $T = D_r(S_r(T))$ on the specified domain, data processing in the reverse direction gives the reverse inequality. No assumption that the answer is deterministically determined by the table is needed. If a serializer adds retrieved facts, removes rows, or changes ambiguous values, the invertibility premise fails and the statement no longer controls the comparison. The proposition therefore motivates both round-trip tests and identical information access across action choices. It does not equate the computational difficulty of extracting the answer from each view.

A.3 Proof of Proposition 2

Condition on the available information $\mathcal{I}$. Returning the primary has conditional objective $-r_0(\mathcal{I})$. Purchasing continuation has objective

$$-r_1(\mathcal{I}) - \lambda c_1(\mathcal{I}) - \mu \ell_1^{\text{time}}(\mathcal{I}) - \eta,$$

where η is the relaxed shadow price of one continuation slot. Continuation is selected exactly when its objective exceeds the primary objective. Rearranging gives Equation (3); equality permits either action or randomization. The result is a one-step relaxation: under a finite horizon, the exact shadow price is generally state-dependent, so the constant η is an approximation rather than a claim about the deployed quota controller.

A.4 Proof of Theorem 1

Expanding the scalar utilities cancels the reference action:

$$u_{a_R} - u_{a_M} = \Delta p_R - \Delta p_M - \lambda(\Delta c_R - \Delta c_M) - \mu(\Delta \ell_R - \Delta \ell_M) \tag{12}$$

$$\ge \underline{g}_R - \overline{g}_M - \lambda(\overline{d}_R - \underline{d}_M) - \mu(\overline{t}_R - \underline{t}_M). \tag{13}$$

Equation (7) makes the final expression positive. For the Pareto claim, $p_{a_R} - p_{a_M} \ge \underline{g}_R - \overline{g}_M \ge 0$, and the analogous cost and latency differences are nonpositive. A strict bound forces at least one strict component difference.

For the efficiency statement, assume $0 < \underline{d}_j \le \Delta c_j \le \overline{d}_j$ and $0 \le \underline{g}_j \le \Delta p_j \le \overline{g}_j$. Then

$$\frac{\Delta p_R}{\Delta c_R} \ge \frac{\underline{g}_R}{\overline{d}_R} > \frac{\overline{g}_M}{\underline{d}_M} \ge \frac{\Delta p_M}{\Delta c_M}. \tag{14}$$

If a representation change reduces cost, an incremental positive-cost ratio is inappropriate; the componentwise dominance test remains well-defined. If estimated confidence intervals are used for the six increments, all bounds must hold simultaneously. For example, a union bound allocates failure probability over all reported intervals and all contexts to which the certificate is claimed to apply. Neither per-example self-confidence nor development point estimates supply this coverage.

For model upgrades, the full alternative cost difference is

$$\Delta c_M = 10^{-6}\big(q^{\text{in}}_{m'} n^{\text{in}}_{m',r} + q^{\text{out}}_{m'} n^{\text{out}}_{m',r} - q^{\text{in}}_{m} n^{\text{in}}_{m,r} - q^{\text{out}}_{m} n^{\text{out}}_{m,r}\big). \tag{15}$$

Different tokenizers or gateway accounting may produce different input counts even for the same string. This expression avoids assuming their equality. Local serialization and scheduling times enter ℓ separately; no hardware-normalized compute claim is inferred from dollar prices.

A.5 Theorem 3 and Proof

Theorem 3 (Joint-routing value after estimation and selection overhead)**.** *Fix a restricted action set $B \subseteq \mathcal{A}$, such as all models at one representation. Let $V_{\mathcal{A}} = \mathbb{E}\max_{a\in\mathcal{A}} u_a(X)$ and $V_B = \mathbb{E}\max_{a\in B} u_a(X)$. If $\widehat{a}$ maximizes estimated utility and $\sup_a |\widehat{u}_a(x) - u_a(x)| \le \epsilon(x)$, then, for nonnegative selection overhead $h(x)$ in utility units,*

$$\mathbb{E}[u_{\widehat{a}(X)}(X) - h(X)] - V_B \ge (V_{\mathcal{A}} - V_B) - 2\mathbb{E}\epsilon(X) - \mathbb{E}h(X). \tag{16}$$

Proof. Let $a^*(x) \in \arg\max_{a\in\mathcal{A}} u_a(x)$. For every x where the assumed approximation bound holds,

$$u_{a^*}(x) - u_{\widehat{a}}(x) = u_{a^*}(x) - \widehat{u}_{a^*}(x) + \widehat{u}_{a^*}(x) - \widehat{u}_{\widehat{a}}(x) + \widehat{u}_{\widehat{a}}(x) - u_{\widehat{a}}(x) \tag{17}$$

$$\le \epsilon(x) + 0 + \epsilon(x). \tag{18}$$

Integrating yields $\mathbb{E}u_{\widehat{a}} \ge V_{\mathcal{A}} - 2\mathbb{E}\epsilon$. Subtracting $\mathbb{E}h$ and V_B gives Equation (16). Separate success, cost and latency bounds imply

$$\epsilon(x) \le \epsilon_p(x) + \lambda\epsilon_c(x) + \mu\epsilon_\ell(x). \tag{19}$$

A hand-set bias contributes to approximation error unless it is itself justified as part of a success estimate. Larger action sets can increase both ideal headroom and estimation error. The sufficient inequality for realized scalar improvement is $V_{\mathcal{A}} - V_B > 2\mathbb{E}\epsilon + \mathbb{E}h$, not merely $|\mathcal{A}| > |B|$. □

Corollary 1 (Restriction gap)**.** *For the best fixed-representation router,*

$$\Delta_{\mathcal{R}} = \mathbb{E}\max_{m,r} u_{m,r}(X) - \max_r \mathbb{E}\max_m u_{m,r}(X) \ge 0. \tag{20}$$

The analogous fixed-model gap is nonnegative. Either can equal zero despite heterogeneous per-action accuracy.

Proof. For every fixed r, pointwise maximization over (m, r') includes all choices at r. Integrate and maximize the restricted value over r. Equality holds, for example, if a single representation is utility-optimal for every context. Differences in average scores alone do not rule that case out. □

A.6 Model–representation interactions

A matched conditional interaction can be written as

$$J(x; m, m', r, r') = p_{m,r}(x) - p_{m,r'}(x) - p_{m',r}(x) + p_{m',r'}(x). \tag{21}$$

If $p_{m,r}(x) = \alpha_m(x) + \beta_r(x)$, this interaction vanishes. Nonzero interaction identifies a failure of that additive success model, but does not by itself guarantee an exploitable routing gain: resource differences, predictor error, and observation noise remain. Conversely, even additive context-dependent format preferences can yield value over a globally fixed format. Interaction and adaptive-format value are therefore distinct empirical questions. A frozen policy's marginal action counts identify neither one.

A.7 Proof of Theorem 2 and accounting scope

Initially $K_0 = 0 = \lfloor \rho \cdot 0 \rfloor$. Assume $K_{i-1} \le \lfloor \rho(i-1) \rfloor$. If $G_i = 0$, then $K_i = K_{i-1} \le \lfloor \rho i \rfloor$. If $G_i = 1$, the gate condition implies $K_{i-1} \le \lfloor \rho i \rfloor - 1$, hence $K_i \le \lfloor \rho i \rfloor$. Each prefix thus satisfies

the bound. Let H_i^0 and H_i^1 be actual HTTP-start indicators. The implementation's one-attempt rule gives $H_i^0 \le 1$ and $H_i^1 \le G_i$, hence

$$\sum_{i=1}^{N}(H_i^0 + H_i^1) \le N + \sum_i G_i \le N + \lfloor \rho N \rfloor. \tag{22}$$

Queue timeouts can reduce starts below the scheduled count. A crashed attempt with unknown outcome must retain its start record and consume its slot if a run supports recovery. Hidden SDK retries would invalidate the premise; the transport is one HTTP attempt per child process. This bound is on attempts, not exact dollars: unknown billed tokens prevent a provider-enforced financial guarantee.

The prefix rule conservatively spreads opportunities through the stream. The first three positions have no slot at $\rho = .25$. It may deny difficult early queries while leaving unused slots later. It is not an optimal stochastic knapsack policy. Its strengths are reproducibility, bounded prefixes, and freedom from test-label feedback. A smarter allocation policy requires a separate prospective evaluation and the same complete cost accounting.

A.8 Theorem 4 and Proof

Theorem 4 (Idealized risk bound with quota denial)**.** *Partition executions into accepted primaries $S = \{Z = 0\}$, denied requests $D = \{Z = 1, G = 0\}$, and granted recovery $G = 1$. If a valid envelope satisfies $r_0(\mathcal{I}) \le \overline{r}_0(\mathcal{I}) \le \gamma$ on S, and conditional final error on granted cases is at most r_f, then*

$$1 - A_\pi \le \gamma \mathbb{P}(S) + \mathbb{P}(D) + r_f \mathbb{P}(G). \tag{23}$$

For the deployed proxy, the weaker assumption $r_0 \le 1 - s + \varepsilon$ on S replaces γ by $\gamma + \varepsilon = 1 - \tau + \varepsilon$. A valid denied-group error bound r_d replaces $\mathbb{P}(D)$ by $r_d \mathbb{P}(D)$.

Proof. Let $\mathcal{H}$ include the gate's score, evidence and permitted predecision information. Since S, D, G partition the run,

$$\begin{aligned}
\mathbb{E}[1 - Y_\pi] &= \mathbb{E}[\mathbf{1}_S(1 - Y_0)] + \mathbb{E}[\mathbf{1}_D(1 - Y_0)] + \mathbb{E}[\mathbf{1}_G(1 - Y_1)] && (24)\\
&= \mathbb{E}[\mathbf{1}_S \mathbb{E}(1 - Y_0 \mid \mathcal{H})] + \mathbb{E}[\mathbf{1}_D(1 - Y_0)] + \mathbb{P}(G)\mathbb{E}[1 - Y_1 \mid G] && (25)\\
&\le \gamma \mathbb{P}(S) + \mathbb{P}(D) + r_f \mathbb{P}(G). && (26)
\end{aligned}$$

For the proxy corollary, substitute $\gamma + \varepsilon = 1 - \tau + \varepsilon$ for the valid accepted-risk envelope. If $\mathbb{P}(G) = 0$, the last term is zero without requiring a defined conditional distribution on an empty event. The same convention applies to D. A denied-group error bound gives the refined expression. A global reliability curve does not automatically validate a conditional bound after evidence selection or a history-dependent quota; calibration must be checked for the relevant selection mechanism. □

A.9 Independent risk audits

A finite audit can bound group-average errors without interpreting every confidence value as a probability. Freeze the policy and three selection groups before collecting an independent audit. If the $n_j > 0$ audited errors in group $j \in \{S, D, G\}$ are independent Bernoulli observations from its deployment-conditional distribution, Hoeffding's inequality and a union bound give, simultaneously with probability at least $1 - \delta$,

$$r_j \le \min\left\{1, \widehat{r}_j + \sqrt{\frac{\log(3/\delta)}{2n_j}}\right\}, \qquad j \in \{S, D, G\}. \tag{27}$$

Here r_G is the final-selection error, whereas r_S, r_D are primary errors (Hoeffding, 1963). With known deployment group masses, their weighted bounds certify total risk by the same decomposition as Theorem 4. Estimated masses require additional uncertainty accounting; an empty group receives the worst-case bound one. Reused questions from a table and quota-coupled observations need not satisfy independence or deployment-conditional sampling. Equation (27) is therefore an audit design, not a numerical certificate for the current run. It also does not establish pointwise conditional calibration. The reported HiTab intervals instead use table-cluster resampling for the paired empirical accuracy difference.

A.10 AN IDEAL RESOURCE-PRICED GATE

Let O be the observable first-response information, and suppose the conditional net correctness gain $d(O) = \mathbb{E}[Y_1 - Y_0 \mid O]$, fallback cost $c_1(O)$ and extra service latency $\ell_1(O)$ are known under a fixed serving regime. For a relaxed randomized gate $g(O) \in [0, 1]$ with $\mathbb{E}g \leq \rho$, the scalar objective relative to the primary is

$$\max_g \mathbb{E}\big[g(O)(d(O) - \lambda c_1(O) - \mu \ell_1(O))\big]. \tag{28}$$

Its Lagrangian introduces $\eta \geq 0$ for the call budget. Pointwise maximization accepts escalation where $d - \lambda c_1 - \mu \ell_1 > \eta$, rejects it where the expression is smaller, and may randomize on equality. When such a gate satisfies feasibility and complementary slackness it solves the relaxed problem. The deployed confidence/evidence gate is an observable approximation; it does not estimate d from held-out labels or claim to solve this optimization. The derivation clarifies that uncertainty alone is insufficient: escalation is valuable where its *net rescue probability* exceeds its resource penalty.

A.11 ANTICIPATING CONTINUATION VALUE

Under the relaxed call price, an ideal primary selector would also anticipate continuation:

$$\begin{aligned} a^*(x) &\in \arg\max_a \{u_a(x) + v_a(x)\}, \\ v_a(x) &= \mathbb{E}[(d_a(O) - \lambda c_{1,a}(O) - \mu \ell_{1,a}(O) - \eta)_+ \mid x, a]. \end{aligned} \tag{29}$$

The implemented primary selector omits this unobserved continuation term. It is consequently a myopic surrogate, not an optimal solution of the sequential problem. If $v_a(x)$ denotes that term, its additional utility gap relative to the relaxed selector is bounded by $\max_a v_a(x) - \min_a v_a(x)$ plus the $2\epsilon(x)$ primary-estimation bound. This follows by adding and subtracting the selected action's primary utility. The expression identifies exactly what must be estimated to improve the controller beyond structural priors.

A.12 CONDITIONAL TOPOLOGY BOUNDARY

For a model m, let $g_r(x) = p_{m,r}(x) - p_{m,\mathrm{CSV}}(x)$ and let d_r, t_r be the corresponding full cost and latency increments. CSV is utility-optimal over that model's formats whenever

$$g_r(x) \leq \lambda d_r(x) + \mu t_r(x) \quad \text{for all } r \in \mathcal{R}. \tag{30}$$

This follows by rearranging each pairwise utility comparison. A topology-dependent boundary would require both resource increments and residual accuracy gains to be measured conditional on topology. Byte inflation alone cannot establish it. Indeed, the frozen diagnostic cohorts have median HTML/CSV byte ratios 2.54 on HiTab and 2.84 on WTQ; these local measurements are not billed token counts and do not support a blanket claim of greater relative HTML overhead on HiTab.

A.13 COUNTEREXAMPLES TO UNCONDITIONAL GUARANTEES

If a fixed decoder ignores markup, all representations have identical correctness; extra HTML tokens can only worsen cost. If HTML gains accuracy but triples token count, it need not dominate a model upgrade. If an escalated model systematically misreads units, increasing escalation can harm accuracy even when it is larger. If a confidence-1.0 answer cites the wrong but internally consistent row, the evidence check passes while semantic correctness fails. Each example obeys the basic architecture and refutes a claim of automatic representation dominance or an unconditional high-accuracy fallback safety net.

B IMPLEMENTATION, DATA, AND MEASUREMENT PROTOCOL

B.1 MODELS, PROMPTS, AND SERIALIZER INVARIANTS

The configured model identifiers are `gemini-3.1-flash-lite`, `gemini-3-flash-preview`, and `gemini-3.1-pro-preview`. The endpoint is an OpenAI-compatible gateway; benchmark claims concern its recorded versions and billing semantics rather than guaranteed equivalence to another provider endpoint. Prices below are configured USD rates, not independently audited invoices.

Table 4: Configured prices per million tokens. Output tokens include reported reasoning-token subsets without double counting.

Model	Input USD	Output USD
Flash-Lite	0.25	1.50
Flash	0.50	3.00
Pro	2.00	12.00

Every serialization begins with shared JSON metadata and includes every physical cell. CSV uses explicit quoting; Markdown escapes pipes, backslashes and line breaks, with synthetic column labels distinguished from content; HTML escapes cell text and records physical coordinates. HiTab merge and hierarchy metadata are identical across views. WTQ uses a flat top-header tree and no inferred left hierarchy. No table title is inferred from the answer: WTQ's context identifier serves as its available table identifier.

The primary and fallback prompts use temperature zero and a maximum of 2,048 output tokens. Prompts request a nonempty answer list, confidence, operation name, and at most 16 grounded citations. Supported operations are lookup, sum, ordered difference, ratio, and percentage; count and other operations remain unverified because a cited set need not be complete. Confidence is requested as the estimated probability of whole-answer correctness. Models can violate the schema or overstate confidence; these are measured outcomes, not grounds to remove examples.

B.2 HiTab provenance and exposure

The 150-example development pilot samples distinct tables and uses all nine action configurations for development diagnostics. It is not a random question sample from the final test set. Leave-one-table-out selection compares representation-only and model-only action subsets with cost weights 0, 5, and 20. The selection utility is accuracy minus five times USD per question. The CSV-only candidate and the joint cheap-model CSV/Markdown candidate both obtain 106/150, and CSV-only is slightly cheaper. The saved full-test selector consequently uses CSV, with 1,107 Flash and 477 Lite primaries.

The separate 30-question diagnostic touches eight test tables. Table 1 reports the full test set; Section 6 also gives accuracy and paired gain after excluding all 33 questions on those tables. After this full run, further tuning against its labels would require a new evaluation split or an explicitly exploratory label. The one malformed source question stored as the singleton list `["2016"]` is unwrapped without inventing a replacement question and remains in every relevant denominator.

B.3 WTQ parsing and versioned official rescoring

The supplied WTQ folder contains 421 escaped CSV files and 4,344 questions in `pristine-unseen-tables.tsv`. A matching tagged file supplies canonical target denotations. The documented CSV dialect uses backslash-escaped quotes and backslashes. All 421 files parse as rectangular tables with that dialect; 23 appear ragged under a default parser. The three serializers round-trip all tables. This verifies content preservation, not model comprehension. The completed archive reports known WTQ cost of $31.616331 and an accounted total of $32.02876075, including $0.41242975 in reserves for 41 unknown bills. Exact billed macro cost remains unavailable.

The archived Python 3 scorer omitted canonical target annotations and differed in numeric parsing and set semantics. We preserve its reports and independently rescore saved predictions using a minimal Python 3 compatibility port of the supplied official evaluator v1.0.2. Matching semantics, normalization, set deduplication and legacy date handling are preserved. Differential tests compare the port with an independently mechanically converted reference on every non-null saved primary and final prediction, as well as explicit numeric, date, normalization and duplicate-answer cases. All comparisons agree. Tagged raw target strings differ from the archived loader's strings on 57 questions; these differences are recorded rather than silently overwritten. Scores from the canonical-target official rescore are the primary WTQ endpoint. Ordered strict accuracy uses the archived reference strings and remains a separate protocol diagnostic.

B.4 Concurrency, timeouts, and durable accounting

The locked configuration permits six workers and two simultaneous requests per key over three credential slots. Eligible keys rotate round-robin; rate-limit responses impose a shared per-key cooldown. The online protocol disables automatic HTTP retries and separate schema-repair calls. Each attempt has a 90-second wall deadline and the logical request, including waiting for a key, has a 110-second deadline. Blocking transport work runs in a subprocess that is terminated and reaped on cancellation or timeout. Tokens and answer text are recorded when available; credentials are not part of the artifact package.

Waves synchronize primary completion before gate grants are committed in the immutable stream order. Consequently, even a non-escalated response may wait for slower primaries in its wave. Primary-only timing is its observed time before that synchronization and shares the cascade run's load. It is not a matched standalone serving measurement; total wall time, service time and request-arrival response time must not be conflated.

If $\mathcal{J}_i$ is the set of started attempts for question i, macro cost is

$$\overline{C} = \frac{1}{N} \sum_i \sum_{j \in \mathcal{J}_i} 10^{-6} (q^{\text{in}}_{m_{ij}} n^{\text{in}}_{ij} + q^{\text{out}}_{m_{ij}} n^{\text{out}}_{ij}). \tag{31}$$

For unknown billing, let $\mathcal{U}$ index missing attempt costs and R_j be their admission reserves. Report known subtotal C_{known} and accounting estimate $C_{\text{acct}} = C_{\text{known}} + \sum_{j \in \mathcal{U}} R_j$. The reserve is neither a measured payment nor a mathematical upper bound on a nonstandard provider invoice. Monetary plots must label estimates and cannot silently treat unknowns as zero.

HiTab records 4,001,720 known input and 1,515,938 known output tokens. The known subtotal is \$10.16819025; accounting adds \$0.067312 to yield \$10.23550225. The corresponding known subtotal per scheduled question is \$0.00641931, distinct from the exact all-attempt macro cost, which is unavailable. Primary known/accounted totals are \$4.96870225/\$4.99260825; fallback known/accounted totals are \$5.199488/\$5.242894.

B.5 Uncertainty, frontier estimands, and missing counterfactuals

Resample table IDs with replacement and retain all their questions, recomputing question-weighted accuracy on each resample. Use the same draw for the primary and cascade to preserve pairing. Percentile intervals summarize 2,000 fixed-seed draws. An unpaired binomial interval would ignore within-table dependence. Baseline differences and frontier intervals require the same table-level pairing wherever comparable outcomes are available.

Let $\mathcal{P}$ be the finite observed policy family. Define normalized hypervolume as the volume of the union of boxes from $(1 - A_\pi, C_\pi / C_{\text{ref}}, L_\pi / L_{\text{ref}})$ to $(1, 1, 1)$, using $C_{\text{ref}} = 0.05$ USD and $L_{\text{ref}} = 120$ seconds. A cost AUC is

$$\frac{1}{C_{\text{ref}}} \int_0^{C_{\text{ref}}} \max\left(\{A_\pi : C_\pi \le b,\ L_\pi \le L_{\text{ref}}\} \cup \{0\}\right) db, \tag{32}$$

with an analogous latency AUC. These are stepwise finite-policy envelopes, not interpolation through untested policies or a unique "three-dimensional AUC." A two-point curve cannot establish improvement over an absent policy family.

A deterministic selector has zero action propensity for unselected actions at a context. Accordingly, its logs do not identify the full fixed-model baselines or the counterfactual best representation. Importance weighting would require overlap and known nonzero propensities (Dudik et al., 2011). Comparing conditional subgroups assigned different actions also confounds action effects with question difficulty. The final matched comparator runs identify their specified contrasts; other unobserved action effects remain unidentified. The final experiment uses matched subsets rather than a full factorial test grid.

C Budgeted Multi-Anchor Final Protocol

The master WTQ cohort samples 180 of 421 tables by stratified simple random sampling, with strata defined by table cell count (≤ 80, 81–300, > 300), and samples up to five questions per selected

Table 5: Completed frozen follow-up allocation. Failed requests and shadow controls count toward the research ledger. A deployed PRISM question still receives at most two calls; its shadow is a separate experimental alternative.

Arm	Questions	Actual starts	Maximum starts
Fresh WTQ PRISM	851	1063	1063
WTQ Pro/CSV	851	851	851
Model-only / CSV primary	572	715	715
Representation-only / Flash primary	572	714	715
Representation swap	380	380	380
Historical rescued cases, repeat	333	333	333
Matched same-model continuation	212	212	212
HiTab paired CSV / HTML	140 pairs	280	280
WTQ Pro/Markdown diagnostic	85	85	85
Total		4633	4634

table. This produces 851 questions. Nested samples of 120, 120, and 20 tables define the one-axis, swap, and Pro/Markdown cohorts, respectively. Sampling treats each table as a cluster, and paired arms intentionally share questions. These WTQ tables are not disjoint from the earlier full WTQ run; no fitting was performed on the new cohort. HiTab samples 100 of 530 unexposed tables and up to two questions each, producing 140 pairs. The seed and all sample IDs, inclusion probabilities, tasks and policy supports are frozen before new inference.

If stratum h has H_h tables, n_h are selected, and table t has N_t questions of which k_t are sampled, each selected question receives weight $w_{ti} = (n_h/H_h)^{-1}(k_t/N_t)^{-1}$. Nested table subsamples multiply by their inverse conditional inclusion probability. Report the weighted ratio $\widehat{A} = \sum w_{ti}Y_{ti}/\sum w_{ti}$ and the unweighted sampled score; do not label either as a complete-benchmark measurement. Both sides of each paired comparison use the narrower cohort's weights. Approximate stratified table-cluster bootstrap intervals omit finite-population corrections and preserve pairing. The three primary accuracy comparisons (PRISM against Pro/CSV and each one-axis control) additionally receive Bonferroni-adjusted bootstrap intervals. Overlapping separate confidence intervals establish neither equivalence nor noninferiority. No noninferiority margin is selected after seeing comparator results.

All 333 historical rescues received one new same-model, same-format repeat after their archived failed primary. This outcome-selected cohort measures conditional recovery, not a deployable policy's accuracy. To avoid selecting only successful Pro outcomes in the prospective control, every newly admitted PRISM fallback also received a same-model shadow continuation. The two alternative policies share the same primary, gate, grant and final-selection rules. Each alternative is charged its primary plus its own continuation; the research ledger charges all actual starts. The shadow answer cannot affect the deployed PRISM answer or subsequent grants. All analyses retain harms and failures as well as rescues.

New arms are interleaved in a frozen schedule under the same endpoint and concurrency. Service latency is recorded per logical request, including key waits, and summed over the stages of an alternative policy; it excludes serialization and unrelated wave synchronization. Shadow and deployed requests occupy different positions in shared serving waves, so their latency contrast is conditional on this mixed experimental load, not a dedicated-service benchmark. These timings can be compared within the new batch but are not substituted for the legacy wave-latency columns. Per-request tokens, reasoning-token subsets, truncation, transport failures and unknown billing are retained. The protocol permits no hidden retries; every continuation is a separately recorded budgeted call. Interrupted execution resumes cached requests and never resubmits an orphaned start.

All final empirical claims use completed measurements from this finite design. Uncalled cells of a nine-action grid are outside the empirical claim. Pro/CSV is a measured high-fidelity anchor, not a statistical upper bound on achievable accuracy. Representation swaps are controlled directional contrasts, not a guarantee that the deliberately changed format loses. The 85-question Pro/Markdown diagnostic extends coordinate coverage without claiming adequate power for a small Pro-format effect. The final cohort was designed after inspecting previous benchmark results; it is a frozen follow-up study, not an untouched replication dataset.

C.1 Complete sampled results and paired uncertainty

The estimates below describe the completed follow-up only. WTQ uses official denotation scoring and HiTab uses strict scoring. Arms with different samples must be compared using matched cohort rows, not their marginal means. The historical row is conditional on the 333 previously rescued outcomes; its resource use is the incremental retry only. All scheduled calls use 2,048-token ceilings.

Table 6: All sampled-arm resource estimates. Cost is accounted mUSD/question (1 mUSD $= \$0.001$); latency is logical request seconds. Brackets are marginal 95% intervals. The final two rows are paired point differences on the common 572-question cohorts and are not differences of marginal rows. Unknown billing uncertainty is not included in these sampling intervals.

Arm	mUSD/q [95% CI]	Seconds [95% CI]
Historical rescue retry	8.597 [8.038, 9.144]	19.24 [17.58, 20.97]
HiTab CSV	3.491 [2.795, 4.238]	11.29 [8.61, 14.61]
HiTab HTML	4.147 [3.345, 5.036]	13.45 [10.09, 17.98]
Model-only	8.774 [7.932, 9.696]	16.93 [15.43, 18.58]
PRISM	8.080 [7.277, 8.941]	15.56 [14.36, 16.82]
Primary	3.419 [3.215, 3.642]	12.31 [11.30, 13.45]
Pro/CSV	12.957 [12.274, 13.660]	12.57 [11.90, 13.26]
Pro/Markdown	13.436 [11.278, 15.675]	12.57 [10.94, 14.30]
Format-only	8.944 [7.873, 10.154]	18.67 [16.91, 20.69]
Same-model continuation	4.927 [4.502, 5.355]	16.83 [15.27, 18.59]
Swap	3.282 [2.978, 3.614]	11.69 [10.26, 13.25]
Matched 572: PRISM − model-only (paired Δ)	-1.039	-1.16
Matched 572: PRISM − format-only (paired Δ)	-0.461	-2.97

C.2 Observed frontier families and diagnostics

Table 7 compares policies on identical questions and inclusion weights within each row. The 851-question family contains primary, same-model continuation, PRISM and Pro/CSV. The Pro/CSV estimate on these questions is 84.89%.

The final audit recomputes normalized hypervolume and constrained finite-policy envelopes as defined in Appendix B. All observed points satisfy both reference ceilings, so the original analyzer's separate cost/latency AUC implementation gives the same point values here. The additional 2,000 stratified cluster draws use seed 20260923 and jointly resample every policy coordinate. Bootstrap uncertainty is conditional on accounted reservations. These are descriptive family summaries, not evidence of superiority to an unmeasured optimal frontier or confidence bands uniform over budgets.

Fresh PRISM requests continuation on 483/851 questions, grants 212 (24.91%), and denies 271. Its weighted mean calls/question is 1.2443, while the raw mean is 1.2491. The deployed final answers contain 104 invalid responses; the same-model alternative contains 132. The model-only arm grants 143/343 requests and records 58 rescues and one harm; format-only grants 142/339 and records 50 rescues and one harm. Their mean accuracies cannot be subtracted from the master PRISM mean because the cohorts differ.

Fresh primary allocation is 314 Lite/HTML, 451 Flash/Markdown and 86 Flash/CSV. The format-only arm selects Flash/Markdown 509 times and Flash/CSV 63 times. Model-only always selects Flash/CSV. The swap's action-conditional weighted chosen-minus-swapped differences are -1.54 points for the 140 Lite/HTML cases, -0.52 for the 198 Flash/Markdown cases, and $+3.66$ for the 42 Flash/CSV cases. These small descriptive subsets do not establish model–representation interaction.

The 85-question Pro format diagnostic has weighted CSV/Markdown accuracy 76.71/77.93%, with paired difference -1.22 points and interval $[-6.10, 3.62]$. On the HiTab pairs, weighted known-cost subtotals are \$0.003361 for CSV and \$0.003335 for HTML, while accounted costs including reserves are \$0.003491 and \$0.004147. Two CSV and nine HTML timeouts make the accounting distinction material. Neither the uncertain accuracy contrast nor these incomplete bills establish the proposed topology mechanism. No known-cost subtotal is treated as an exact invoice with unknown requests assigned zero cost.

Table 7: Exploratory finite-family frontier summaries with joint table-bootstrap 95% intervals. References are $0.05/question and 120 seconds.

Cohort	Hypervol.	Cost AUC	Lat. AUC
851 q.	0.690	0.769	0.761
95% CI	[0.666, 0.714]	[0.744, 0.794]	[0.737, 0.786]

Table 8: WTQ scorer reconciliation on identical saved predictions. Official denotation uses canonical targets.

Policy	Strict	Arch.	Offic.	Calls/q
Primary	67.24	69.52	70.72	1.000
PRISM	73.64	76.52	78.04	1.250

D Gate Diagnostics and Completed Ablations

Table 9: Exact locked HiTab diagnostic counts after reconciling one reason-ledger rescue to the official rescore. No observed confidence was below 0.8.

Reason	Quest.	Grant.	Den.	Resc.	Harm
Consistent evidence	1164	0	0	0	0
Calculation mismatch	119	101	18	28	8
Incomplete response	13	13	0	4	0
Invalid answer	101	90	11	37	0
Invalid coordinates	1	1	0	1	0
Missing/excess evidence	4	3	1	0	0
Unsupported check	36	34	2	8	0
Ungrounded/duplicate	144	132	12	4	3
Transport failure	2	2	0	1	0
Total	1584	376	44	83	11

Table 10: Paired correctness confusion matrices for HiTab (strict) and WTQ (denotation).

Dataset	Primary	Cas. wrong	Cas. corr.
HiTab	Wrong	258	83
	Correct	11	1232
WTQ	Wrong	939	333
	Correct	15	3057

Table 11: WTQ gate diagnostics after reconciling three reason-ledger rescues to the official rescore. These observed subgroups are not randomized policy ablations; the reconciled aggregate totals match the cohort tables.

Reason	Questions	Granted	Denied	Rescues	Harms
Calculation mismatch	255	114	141	95	10
Consistent evidence	1988	0	0	0	0
Incomplete response	37	19	18	15	0
Invalid answer	746	342	404	150	0
Missing/excess evidence	35	19	16	0	1
Unsupported semantic check	1195	555	640	61	4
Transport failure	41	14	27	12	0
Ungrounded/duplicate evidence	47	22	25	0	0
Reason-ledger subtotal (official-rescore reconciled)	4344	1085	1271	333	15

WTQ primary confidence values are 1.0 on 3,436 examples, 0.9 on 70, 0.8 on seven, 0.7 on one, 0.5 on three, and 0.0 on three; 824 are missing. Only seven finite scores fall below 0.8. Of 1,988 accepted consistent-evidence primaries, 179 are incorrect under official denotation scoring. The fallback rescues 30.69% and harms 1.38% of granted cases; its conditional net gain is 29.31%. The paired confusion counts are 939 wrong/wrong, 333 wrong/correct, 15 correct/wrong, and 3,057 correct/correct.

D.1 Ablation scope and counterfactual support

The completed no-cascade ablation loses 4.55 strict points on HiTab and 7.32 official-denotation points on WTQ. The final experiment completed the matched one-axis, format-swap, and same-model continuation comparisons in Table 5. Broader confidence/quota sweeps, alternative-provider replication and factorized learned routers are not measured by this budget and are not presented as completed results. A retrospective threshold or quota replay is valid only when every needed continuation outcome is observed. Neither a passed evidence check nor the concentrated confidence distribution establishes conditional calibration. The original 0.8 threshold and 0.25 quota are fixed for all matched cascade controls.

Table 12: WTQ gate diagnostics after reconciling three reason-ledger rescues to the official rescore. These observed subgroups are not randomized policy ablations; the reconciled aggregate totals match the cohort tables.

Reason	Questions	Granted	Denied	Rescues	Harms
Calculation mismatch	255	114	141	95	10
Consistent evidence	1988	0	0	0	0
Incomplete response	37	19	18	15	0
Invalid answer	746	342	404	150	0
Missing/excess evidence	35	19	16	0	1
Unsupported semantic check	1195	555	640	61	4
Transport failure	41	14	27	12	0
Ungrounded/duplicate evidence	47	22	25	0	0
Reason-ledger subtotal (official-rescore reconciled)	4344	1085	1271	333	15

E Frozen Policy Parameters and Algorithm

The context partition uses cell counts ≤ 80, 81–300, and > 300. Arithmetic intent is flagged by the words sum, total, average, mean, difference, percent, percentage, ratio, more, less, increase, decrease, or combined. This inexpensive lexical feature is incomplete: for example, a count question need not contain any of those words. Header depth is recorded but does not determine the current six-way context bucket. No gold operation annotations enter the selector.

Table 13: WTQ frozen synthetic priors and global preferences. The k, n notation follows the schema but does not denote measured WTQ outcomes. Configured cost weight is zero.

Model	Format	k	n	Bias	Cost proxy USD
Lite	CSV	78	100	0	.0008
Lite	MD	81	100	.010	.0009
Lite	HTML	83	100	.035	.0010
Flash	CSV	84	100	.010	.0022
Flash	MD	86	100	.035	.0024
Flash	HTML	80	100	0	.0025

Table 14: Additional WTQ context preferences. All other context/action offsets are zero. These are prospective design choices.

Context	Favored action	Offset
Small / lookup	Lite / HTML	.060
Medium / lookup	Flash / MD	.045
Large / lookup	Lite / HTML	.055
Small / arithmetic	Flash / CSV	.050
Medium / arith.	Flash / MD	.050
Large / arithmetic	Flash / CSV	.060

Tie-breaking chooses lower proxy cost and then lexicographic action ID. The WTQ fallback is Pro/HTML; the locked HiTab fallback is Pro/CSV. Both use $\tau = .8$ and $\rho = .25$. The execution caps were 1,980 and 5,430 HTTP starts for complete HiTab and WTQ schedules, respectively. The scoring rule, not a quota over representation labels, yields the displayed format allocation. An imposed preference is transparent; it is not calibrated evidence of superiority.

Figure 2(b) visualizes the same frozen scores in the compact model–representation grid, while Figure 2(a) shows the realized primary allocation after context preferences and tie-breaking are applied.

F Evidence-Conditioned Cohort and Utility Audit

F.1 Retrospective design and interpretation

This analysis was conducted after inspecting the aggregate follow-up results and was not preregistered with the live experiment. Before computing subgroup outcomes, we fixed the cohort definitions in a versioned manifest using the original size thresholds, header depth, lexical intent and gate observations. All 18 named groups are retained for each run, including empty and unfavorable groups. Groups overlap: depth, size and confidence are different descriptors, and admitted cases are nested within requested cases. The artifact tables retain all matched policy contrasts across the input-defined cohorts; no threshold search or policy update was performed.

Large means more than 300 physical cells; nested and deep mean supplied header depth at least two and three. Arithmetic is the original lexical flag, not a verified operation count. WTQ has no

Algorithm 1: Bounded Joint Execution

Input: immutable question stream, frozen action scores, τ, ρ; initialize $K = 0$.

1. For each wave, select one action per input using Equation (6), record the plan, serialize the entire table, and issue one primary attempt.

2. In stream order, inspect each primary response. Set $Z = 1$ for an invalid answer, missing/low confidence, or inconsistent/unsupported evidence.

3. Grant $G = 1$ only if $Z = 1$ and $K < \lfloor \rho i \rfloor$; durably record the gate, then update $K \leftarrow K + G$.

4. If granted, issue at most one fallback attempt. Apply the frozen final-selection rule; do not consult the reference answer.

5. Record prediction, all starts, usage, unknown billing, and timing. After execution, score references offline and package the run.

Figure 3: The separation between inference objects and references prevents online use of test correctness. Concurrent API calls do not change gate admission order.

supplied deep hierarchy. Borderline means $.8 \leq s < .95$, separate from the original trigger $s < .8$; its definition is descriptive and does not imply that it caused an escalation. No multi-hop labels are inferred from incorrectness, fallback success or question phrasing. Available-but-flagged and unavailable-output groups partition requested cases, not epistemic and aleatoric uncertainty.

All cohort intervals are exploratory pointwise 95% intervals from 2,000 table-cluster resamples with seed 20260924. The full sampling frame is resampled, retaining tables with zero subgroup contribution; a subgroup statistic divides its weighted successes by its weighted membership in each draw. Draws with zero membership are recorded as undefined and omitted from percentiles. Final samples retain stratification and inclusion weights. These intervals are not adjusted for the exploratory cohort family, do not re-execute the quota, and can be degenerate for sparse groups. They therefore cannot certify a discovered optimal cohort. Earlier prespecified aggregate intervals remain unchanged.

Table 15: Original full WTQ: complete cohort outcomes. Accuracy uses the run-specific endpoint; R/H gives raw rescues/harms. Empty groups are explicitly marked, not treated as zero accuracy.

Cohort	N	Primary %	Final %	Gain pp [95% CI]	R/H
All	4344	70.72	78.04	7.32 [6.47, 8.21]	333/15
Small ($\leq$ 80 cells)	1659	81.86	84.93	3.07 [2.07, 4.12]	61/10
Medium (81–300 cells)	2338	63.86	74.47	10.61 [9.41, 11.86]	252/4
Large ($>$ 300 cells)	347	63.69	69.16	5.48 [2.74, 8.17]	20/1
Arithmetic lexical flag	839	65.55	74.85	9.30 [7.31, 11.24]	79/1
Large and arithmetic	65	56.92	69.23	12.31 [4.62, 20.99]	8/0
Header depth ≥ 2	0	*Empty cohort*			
Header depth ≥ 3	0	*Empty cohort*			
Confidence $.8 \leq s < .95$	77	53.25	62.34	9.09 [3.33, 15.59]	7/0
Confidence $s < .8$	7	28.57	14.29	-14.29 [-50.00, 0.00]	0/1
Confidence unavailable	824	0.00	21.48	21.48 [18.69, 24.35]	177/0
Requested recovery	2356	53.61	67.11	13.50 [12.02, 15.00]	333/15
Admitted recovery	1085	53.92	83.23	29.31 [26.31, 32.36]	333/15
Denied recovery	1271	53.34	53.34	0.00 [0.00, 0.00]	0/0
Accepted primary	1988	91.00	91.00	0.00 [0.00, 0.00]	0/0
Unavailable answer	824	0.00	21.48	21.48 [18.69, 24.35]	177/0
Available, flagged answer	1532	82.44	91.65	9.20 [7.80, 10.62]	156/15
Large and requested	179	44.69	55.31	10.61 [5.53, 15.19]	20/1

The conditional return on an admitted slot is distinct from cohort affordability: original WTQ admitted cases cost \$0.020495 each, and fresh admitted cases cost \$0.023549, including their primaries. The original whole-run \$0.007373 average cannot be attached to this selected high-recovery cohort. Relative gains are undefined when primary accuracy is zero; the unavailable-output rows therefore report percentage-point gains only.

Table 16: Fresh sampled WTQ: complete cohort outcomes. Accuracy uses the run-specific endpoint; R/H gives raw rescues/harms. Empty groups are explicitly marked, not treated as zero accuracy.

Cohort	N	Primary %	Final %	Gain pp [95% CI]	R/H
All	851	68.19	75.93	7.74 [6.03, 9.58]	73/7
Small ($\leq$ 80 cells)	320	79.74	81.72	1.98 [-0.37, 4.61]	10/4
Medium (81–300 cells)	451	61.12	72.51	11.39 [8.90, 13.93]	55/1
Large ($>$ 300 cells)	80	63.21	72.37	9.16 [2.81, 16.57]	8/2
Arithmetic lexical flag	170	63.89	70.11	6.22 [2.08, 10.73]	11/1
Large and arithmetic	14	48.72	58.97	10.26 [0.00, 37.51]	1/0
Header depth ≥ 2	0	*Empty cohort*			
Header depth ≥ 3	0	*Empty cohort*			
Confidence $.8 \leq s < .95$	13	27.66	18.05	-9.61 [-30.71, 0.00]	0/1
Confidence $s < .8$	1	0.00	0.00	0.00 [0.00, 0.00]	0/0
Confidence unavailable	182	0.00	32.98	32.98 [27.23, 38.55]	61/0
Requested recovery	483	50.24	63.91	13.67 [10.79, 16.69]	73/7
Admitted recovery	212	49.77	81.44	31.67 [25.18, 38.34]	73/7
Denied recovery	271	50.60	50.60	0.00 [0.00, 0.00]	0/0
Accepted primary	368	91.61	91.61	0.00 [0.00, 0.00]	0/0
Unavailable answer	182	0.00	32.98	32.98 [27.23, 38.55]	61/0
Available, flagged answer	301	81.69	83.27	1.58 [-1.01, 4.31]	12/7
Large and requested	50	46.23	60.83	14.60 [4.51, 27.37]	8/2

Table 17: Original full HiTab: complete cohort outcomes. Accuracy uses the run-specific endpoint; R/H gives raw rescues/harms. Empty groups are explicitly marked, not treated as zero accuracy.

Cohort	N	Primary %	Final %	Gain pp [95% CI]	R/H
All	1584	78.47	83.02	4.55 [3.11, 5.99]	83/11
Small ($\leq$ 80 cells)	428	81.78	85.98	4.21 [1.92, 6.83]	20/2
Medium (81–300 cells)	894	77.40	81.43	4.03 [2.11, 5.98]	45/9
Large ($>$ 300 cells)	262	76.72	83.59	6.87 [3.32, 10.19]	18/0
Arithmetic lexical flag	856	77.57	81.31	3.74 [1.83, 5.69]	39/7
Large and arithmetic	107	82.24	91.59	9.35 [3.08, 14.52]	10/0
Header depth ≥ 2	1549	78.37	82.96	4.58 [3.15, 6.05]	82/11
Header depth ≥ 3	838	76.01	80.67	4.65 [2.88, 6.67]	43/4
Confidence $.8 \leq s < .95$	40	62.50	72.50	10.00 [-7.14, 25.46]	7/3
Confidence $s < .8$	0	*Empty cohort*			
Confidence unavailable	116	0.00	36.21	36.21 [27.42, 45.53]	42/0
Requested recovery	420	42.14	59.29	17.14 [11.74, 22.50]	83/11
Admitted recovery	376	43.35	62.50	19.15 [13.25, 25.13]	83/11
Denied recovery	44	31.82	31.82	0.00 [0.00, 0.00]	0/0
Accepted primary	1164	91.58	91.58	0.00 [0.00, 0.00]	0/0
Unavailable answer	116	0.00	36.21	36.21 [27.42, 45.53]	42/0
Available, flagged answer	304	58.22	68.09	9.87 [4.55, 15.31]	41/11
Large and requested	58	24.14	55.17	31.03 [15.38, 47.83]	18/0

Table 18: Resource use within the highlighted cohorts. Accounted mUSD/question includes all primary and admitted continuation costs, not just successful rescues. Original runs use wave-based timing; fresh WTQ uses logical request timing. They cannot be compared as a cross-run speedup.

Cohort	Primary mUSD	Final mUSD	Primary s	Final s
WTQ: requested	4.157	11.699	12.66	32.73
WTQ: admitted	4.118	20.495	12.11	36.39
Fresh WTQ: admitted	4.472	23.549	16.63	29.91
WTQ: large	4.873	15.418	7.17	28.98
Fresh WTQ: large	5.302	19.393	14.15	18.21
HiTab: deep	3.053	6.892	5.70	14.87

F.2 Matched structural controls and heterogeneity

On their respective matched cohorts ($n = 572$), the paired accuracy difference for PRISM over model-only is $+3.85\%$ (95% CI $[0.72, 6.98]$) and over format-only is $+4.20\%$ (95% CI $[1.05, 7.35]$). The archived matched analysis stores these paired estimands and resource summaries, not separate marginal accuracy estimators for each restricted-selector arm; absolute baseline percentages are therefore not reconstructed from a paired difference. Point estimates favor PRISM on all three objectives in each comparison, with PRISM establishing a statistically significant accuracy advantage over both baselines. Paired cost/latency differences are $-\$0.001039/-1.16$ seconds against model-only and $-\$0.000461/-2.97$ seconds against format-only; Appendix C includes their uncertainty.

Across the 380 questions evaluated, the format swap produces a paired accuracy difference of $+2.11\%$ (95% CI $[-1.34, 5.56]$) in favor of the chosen primary. Individual outcomes are sensitive to the two executions, but without same-format replicates those disagreements cannot isolate serialization from generation variability. On HiTab, HTML has a higher accuracy point estimate (82.56% versus CSV's 78.85%) and a wide paired interval. CSV's lower accounted cost is not evidence that hierarchy induces accuracy dominance; its cost comparison is also sensitive to timeout reserves. The proposed topology boundary remains a conditional utility criterion, not a discovered cross-domain law.

For the shared-primary continuation comparison, the large-table advantage is 11.15 points versus 3.15 on other tables: the difference is 7.99 points $[1.27, 15.41]$, an exploratory interaction without multiplicity control. Restricted-selector interactions include zero. The small large-and-arithmetic intersections are retained in the machine-readable audit: PRISM is worse than model-only by 30.30 points on seven questions and worse than format-only by 13.33 points on six. Deep HiTab CSV-minus-HTML is -3.45 points $[-20.35, 9.05]$ on 55 questions. These results preclude interpreting structural complexity as a universal guarantee of joint-policy superiority.

F.3 Utility definitions and the complete comparison family

Define cost productivity $\mathcal{E}_\pi = A_\pi / C_\pi$ for $C_\pi > 0$. Under repeated draws from the same population and resource regime, this is the ratio of expected correct returns to expected expenditure. It is not the average per-question ratio, a minimum-accuracy guarantee, or a Pareto ordering. Latency and output quality remain separate objectives.

For a policy used alone, the budget-integrated accuracy at reference B is

$$\mathrm{AUC}_C^{\mathrm{alone}}(\pi; B) = B^{-1} \int_0^B A_\pi \mathbf{1}\{C_\pi \le b,\ L_\pi \le 120\}\, db = A_\pi (1 - C_\pi/B)_+ \mathbf{1}\{L_\pi \le 120\}. \tag{33}$$

This treats a budget below mean policy cost as infeasible and assigns zero there; it does not invent a randomized low-budget execution. The observed-family AUC in Table 7 instead integrates the best feasible member. The fixed $B = \$0.05$ comes from the original protocol; utility weights are sensitivity parameters, not fitted evidence of optimality.

Table 19: Paired accuracy contrasts (left minus right), using comparison cohort weights. Intervals are marginal 95%.

Contrast	Qs	Diff (pp)	95% CI
PRISM − model-only	572	+3.85	[0.72, 6.98]
PRISM − format-only	572	+4.20	[1.05, 7.35]
Chosen − format swap	380	+2.11	[−1.34, 5.56]

Table 20: Policies on common support: cost productivity and policy-alone AUC ($B = \$0.05$).

Policy	Corr./USD [95% CI]	Cost AUC [95% CI]
Primary	199.45 [180.0, 220.2]	0.6353 [.601, .668]
Same-contin.	146.60 [131.2, 164.0]	0.6511 [.617, .684]
PRISM	93.97 [83.5, 105.8]	0.6366 [.605, .667]
Pro/CSV	65.52 [60.8, 70.1]	0.6289 [.603, .653]

At the 851-question point estimates, the cost-only preference changes from Pro/CSV to same-model continuation and then to primary as λ increases, with ties at the boundaries. PRISM occupies an intermediate deployment point whose value depends on the joint weighting of accuracy, cost, latency, and quota admission; these intervals extend beyond the plotted range and follow from all pairwise line intersections.

The policy-alone AUC difference PRISM minus Pro is .0077 $[-.0154, .0301]$ on 851 questions. This does not support decisive dominance. The stronger supported comparative resource statement

is the 43.43% improvement in accuracy per accounted dollar on the master cohort, alongside 37.64% lower cost, lower accuracy and higher logical latency. No tenfold saving, universal high-capacity latency penalty or near-equivalence in accuracy is inferred. The dollar range from the original full runs cannot replace costs from these matched samples.

G Artifact Audit and Empirical Claim Boundaries

The original HiTab archive is identified by run suffix `20260920_195611_383371Z`; the earlier audit verified 48 hashed artifacts. The WTQ archive is `20260921_074509_485690Z`; its 41 hashes and saved-output official rescore were checked independently. The final follow-up archive is `20260922T120627_489830Z`. All 59 manifest-listed files, all 27 frozen code fingerprints, and the design manifest hash pass verification. Replaying the archived analyzer reproduces metrics, per-question rows, paired contrasts, intervals, gate diagnostics, frontier summaries and analysis summaries exactly. Independent weighted-mean and ledger calculations agree. Scoring every final WTQ policy row against a separately syntax-converted official evaluator also agrees. All 1,255 unique frozen inputs pass CSV, Markdown and HTML round trips (3,765 checks).

The final ledger contains 4,254 unique completed tasks and 4,633 starts, each with one matching terminal attempt record and one logical request record. All attempts have index one. Every cascade prefix respects its 25% cap, and every PRISM grant has exactly one experimental shadow continuation. There are timeout attempts and no other terminal transport error class. These observations support bounded termination for this batch, not the elimination of API failures. Total wall time is 32,019.69 seconds (8.89 hours). The known cost is $36.6664045, reserves add $1.6209795 for unknown bills, and total accounted research cost is $38.287384. The experimental ledger includes controls and cannot be interpreted as deployed PRISM cost.

The reproduced analysis uses the frozen 2,000-draw bootstrap seed. Additional descriptive primary/cascade contrasts and common-support frontier summaries were computed offline without changing predictions or selecting new test examples. The latter use a separate disclosed seed and are not part of the three predeclared accuracy comparisons. Implementation tests pass, including matching semantics, accounting, sampling, bounded execution and interrupted-run behavior. Static source checks do not certify pagination or rendering; final pagination depends on the supplied ICLR style and the two referenced figure assets.

The evidence supports a bounded cascade recovery claim and a resource trade-off against the measured Pro anchor. It does not establish calibrated risk, a topology-induced optimal representation, unrestricted nine-way Bayesian selection, or PRISM Pareto dominance. The later cohort and utility audit verifies all source archives, reproduces every logged gate with Equation (4), and retains every defined cohort. These retrospective results do not change the frozen aggregate comparisons. Remaining unmeasured controls delimit the scope of the paper; none are represented by a fabricated numerical result.