QUOTA: HEAD-SENSITIVE UNCERTAINTY FOR BUDGET-INTERPRETABLE TOKEN ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Inference-time compute scaling underpins state-of-the-art reasoning in large language models (LLMs), but fully autoregressive generation with large models carries prohibitive computational costs. While speculative decoding and collaborative inference mitigate this via small–large model collaboration, existing approaches suffer from coarse routing granularity, reliance on supervised training, or model-dependent uncertainty thresholds that lack interpretable resource semantics. To address this challenge, we introduce QUOTA, a training-free token-level routing framework that decouples where to deploy LLM compute from how much capacity to allocate. QUOTA ranks generation states using head-sensitive collision entropy and calibrates raw scores into percentiles via a prompt-only empirical cumulative distribution function (ECDF). Given a user-specified LLM token quota q, states in the top-q uncertainty tail receive one LLM token, after which routing is recomputed from the updated shared prefix, eliminating auxiliary training or routing labels. Evaluated across mathematical reasoning, scientific QA, and code generation tasks, QUOTA achieves superior quality-latency tradeoffs and precise budget fidelity: with the Qwen3-0.6B/32B model pair, it obtains a 48.55 macro-average pass@1, outperforming SpecReason by 2.90 percentage points with a 17-second reduction in per-question latency; on AIME25 with Qwen3-0.6B/8B, it delivers 42.50% accuracy using only 29% LLM output tokens, surpassing GlimpRouter and SpecReason under lower compute budgets. Across three model pairs and twelve operating points, QUOTA maintains a budget mean absolute error of 1.32 percentage points, offering fine-grained and auditable token-level collaborative inference for efficient high-quality LLM reasoning.

1 INTRODUCTION

Inference-time compute scaling strengthens large language model (LLM) reasoning (Guo et al., 2025; Snell et al., 2025; Wu et al., 2025; Lightman et al., 2024), but long autoregressive trajectories become prohibitively expensive at scale (Kwon et al., 2023; Zhong et al., 2024; Agrawal et al., 2024). Small-large collaboration reduces this cost by delegating routine generation to a small language model (SLM) and invoking the LLM only where needed (Chen et al., 2024; Ong et al., 2025; Zhang et al., 2024b; Ding et al., 2025). Yet reasoning difficulty is non-uniform and model choice matters mainly at sparse, consequential positions. Query-level routing cannot adapt during generation; step-level routing commits an entire span from one local decision. Token-level routing offers finer control, but its core challenge is deciding when and where to invoke the LLM and how much capacity to allocate, requiring a lightweight local-risk signal and a parameter with interpretable LLM usage across models and workloads.

Existing methods address only parts of this problem. Query-level routers fix one model for the response (Chen et al., 2024; Ong et al., 2025; Zhang et al., 2024b; Feng et al., 2025). Step-level routers such as SpecReason and GlimpRouter adapt within a trajectory but still assign full reasoning spans from local decisions (Pan et al., 2025; Zeng et al., 2026). Related methods guide or accelerate reasoning through speculative exploration, reward-guided decoding, process verification, uncertainty-aware branching, or model collaboration (Shi et al., 2025; Liao et al., 2025; Setlur et al., 2025; Li et al., 2026; She et al., 2025a). Learned token-level methods such as Co-LLM, CITER, and R2R enable finer interleaving but require trained policies, paired outputs, rollouts, or divergence labels (Shen et al., 2024; Zheng et al., 2025; Fu et al., 2025). Speculative and parallel decoding

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.

act at token granularity but optimize target-model acceleration rather than selective semantic intervention (Leviathan et al., 2023; Cai et al., 2024; Li et al., 2024; Miao et al., 2024). Training-free adaptive reasoning methods use predictive entropy to allocate computation (Zeng et al., 2026; Li et al., 2026), while semantic uncertainty and confidence calibration are widely studied for generation reliability (Kuhn et al., 2023; Farquhar et al., 2024; Geng et al., 2024). High-entropy tokens also tend to mark consequential reasoning forks (Wang et al., 2025). However, Shannon entropy changes with low-probability tail fragmentation even when leading-candidate competition is stable, and raw thresholds drift across models and workloads. A gap therefore remains for training-free token routing with sharp local selectivity and explicit budget semantics.

To address this gap, we propose **QUOTA**, a training-free token router that separates local risk ranking from budget control. Its head-sensitive collision entropy squares next-token probabilities, emphasizing plausible leading candidates while suppressing diffuse tail mass. A prompt-only empirical cumulative distribution function (ECDF) maps model-specific scores to reference percentiles using only unlabeled prompts and SLM-generated distributions, requiring no correctness or routing labels, paired model outputs, or LLM queries during calibration. The user specifies an LLM token quota q; top-q uncertainty states receive one LLM token before rerouting from the updated prefix. Collision entropy determines where strong-model capacity is spent, while ECDF calibration controls how much, yielding one interpretable parameter across model pairs and tasks. With Qwen3-0.6B/32B, QUOTA reaches the quality region of standalone Qwen3-8B (48.55 versus 48.68 macro-average pass@1) while reducing mean latency from 182 to 167 seconds per question.

Our contributions are summarized as follows.

- **Analysis-Driven Formulation.** We decompose token-level collaboration into prioritizing consequential states and controlling LLM allocation with an interpretable budget, motivating head-sensitive ranking and percentile calibration.
- **QUOTA Framework.** QUOTA combines collision-entropy ranking, prompt-only ECDF calibration, and immediate one-token LLM takeover without router training or routing labels, exposing a quota with direct resource semantics.
- **Empirical Validation.** Across mathematics, scientific QA, code generation, and multiple model pairs, QUOTA improves quality–compute tradeoffs and budget adherence. With Qwen3-0.6B/32B, it reaches 48.55 macro-average pass@1, 2.90 points above SpecReason with 17 seconds per question lower latency; across three pairs and 12 operating points, requested and realized usage differ by only 1.32 points in mean absolute error.

2 Related Work

Our work sits at the intersection of four research threads: collaborative inference across model sizes, token-level compute routing, uncertainty-driven adaptive computation, and confidence calibration. While prior work has separately established the merits of dynamic compute allocation, QUOTA unifies three properties that no single existing method simultaneously achieves: per-token routing granularity, training-free uncertainty ranking, and an interpretable, budget-explicit control interface.

2.1 Collaborative Inference Across Routing Granularities

Collaborative inference systems use heterogeneous model portfolios with varying capability-cost profiles to improve the efficiency-quality trade-off (Chen et al., 2024; Ding et al., 2024; Ong et al., 2025; Zhang et al., 2024b; Ding et al., 2025; Feng et al., 2025). Query-level cascades and preference routers select a single model for an entire response based on input confidence, cost constraints, or learned routing policies; they adapt across requests but cannot reallocate compute once a generation trajectory begins.

Step-level methods refine this allocation by redistributing compute within a single reasoning trajectory. SpecReason interleaves SLM-generated proposal steps with LLM assessment or regeneration (Pan et al., 2025); GlimpRouter dispatches full reasoning steps based on the entropy of the initial token (Zeng et al., 2026); and related systems use speculative exploration, reward-guided decoding, process verification, uncertainty-aware branching, or model collaboration (Shi et al., 2025; Liao et al., 2025; Setlur et al., 2025; Li et al., 2026; She et al., 2025a). Collectively, these approaches

operate over proposal, verification, or reasoning units rather than one-token model selection. In contrast, QUOTA re-evaluates the routing decision after every emitted token, treating strong-model compute as a fine-grained local resource. This allows intervention at arbitrary positions where consequential reasoning forks emerge, even when they do not align with natural step boundaries.

2.2 Learned Token Routing and Speculative Generation

Learned token-level collaboration achieves even finer-grained control over model interleaving. Co-LLM (Shen et al., 2024) dynamically interleaves language models during decoding; CITER (Zheng et al., 2025) optimizes a quality- and cost-aware token-routing policy via reinforcement learning; R2R (Fu et al., 2025) learns to identify tokens at which reasoning paths diverge; and token routing has also been explored in edge-cloud deployment scenarios (She et al., 2025b). While these works demonstrate that sparse intervention can preserve strong-model reasoning quality, they typically rely on trained routers or require supervision from paired model outputs, generation rollouts, divergence labels, or dedicated verifiers. QUOTA, by contrast, derives routing signals directly from the SLM's native next-token distribution and calibrates them using only unlabeled prompts.

Speculative and parallel decoding also operate at token granularity, but pursue a fundamentally different objective: they accelerate target-model generation through proposal-and-verification mechanisms, auxiliary decoding heads, retrieval-based drafts, or multi-token lookahead generation (Leviathan et al., 2023; Cai et al., 2024; Li et al., 2024; He et al., 2024; Fu et al., 2024). The core control problem in speculative decoding is draft–target equivalence and acceptance-rate optimization. QUOTA addresses a distinct problem: selective allocation of LLM capacity to maximize reasoning quality under a compute budget. Rather than accepting or rejecting draft tokens en bloc, QUOTA directly assigns the next visible token to the LLM at high-uncertainty positions and immediately resumes SLM generation from the updated prefix.

2.3 Uncertainty Signals for Adaptive Reasoning

Token-level predictive entropy, computed directly from model logits, provides a natural, training-free signal for adaptive compute allocation (Zeng et al., 2026; Li et al., 2026). Related work studies semantic uncertainty and confidence calibration for generation reliability (Kuhn et al., 2023; Farquhar et al., 2024; Huang et al., 2024; Geng et al., 2024). Shannon entropy is widely used in reasoning systems and collaborative routing (Zeng et al., 2026; Li et al., 2026; Geng et al., 2024), and recent analyses show that high-entropy tokens tend to coincide with consequential reasoning forks (Wang et al., 2025). GlimpRouter, for example, uses initial-token entropy to trigger full reasoning-step dispatches (Zeng et al., 2026).

However, Shannon entropy measures overall distribution spread and is inherently sensitive to fragmentation in the low-probability tail of the distribution, even when competition among the top candidates remains unchanged. For per-token intervention strategies like QUOTA, what matters is head-to-head competition among plausible next-token candidates. To capture this property, QUOTA employs collision entropy as a head-sensitive uncertainty score. By squaring next-token probabilities, collision entropy amplifies differences among leading candidates while attenuating contributions from diffuse, low-probability tail mass.

2.4 Calibration and Budget-Interpretable Routing

Confidence calibration techniques map raw model scores to empirical reliability estimates, enabling consistent comparison across tasks and model checkpoints (Huang et al., 2024; Ulmer et al., 2024; Zhang et al., 2024a; Manggala et al., 2025; Geng et al., 2024). In collaborative routing, however, a central challenge remains: translating an uncertainty threshold into predictable strong-model resource usage. Raw uncertainty scores lie on model- and workload-specific scales, so a fixed numerical threshold conflates risk ranking with score magnitude and yields inconsistent LLM usage rates across settings.

QUOTA addresses this with a prompt-only ECDF calibration step. The ECDF transforms raw collision entropy values into percentiles relative to a reference distribution generated from unlabeled prompts using only the SLM. This preserves the relative ordering of states while expressing each position in normalized percentile units. The resulting percentile threshold has direct tail-mass se-

mantics: it specifies the nominal fraction of tokens assigned to the LLM, defining an explicit, interpretable compute budget. This stands in contrast to task-specific confidence scores, preference weights, verification thresholds, or learned router parameters.

3 METHODOLOGY

3.1 OVERVIEW

QUOTA operates in two phases: an offline prompt-only rank calibration phase and an online budget-interpretable token routing phase. During calibration, the SLM autoregressively processes unlabeled prompts and constructs a frozen ECDF over token-level collision entropy values. During inference, every shared prefix yields a collision entropy score, which is mapped to a calibrated percentile via the precomputed ECDF. A user-specified quota q determines whether the SLM or LLM emits the next token, and the routing decision is recomputed after every token emission. The two components serve distinct, complementary roles: collision entropy identifies where strong-model capacity is most likely to improve reasoning, while ECDF calibration converts that ranking into an explicit, predictable allocation budget.

Figure 1 summarizes the end-to-end pipeline. Routing decisions depend solely on the SLM's native next-token distribution and the frozen calibration artifact. QUOTA requires neither auxiliary router training nor an LLM forward pass to score the current state; the LLM is invoked exclusively for tokens falling within the requested high-risk tail.

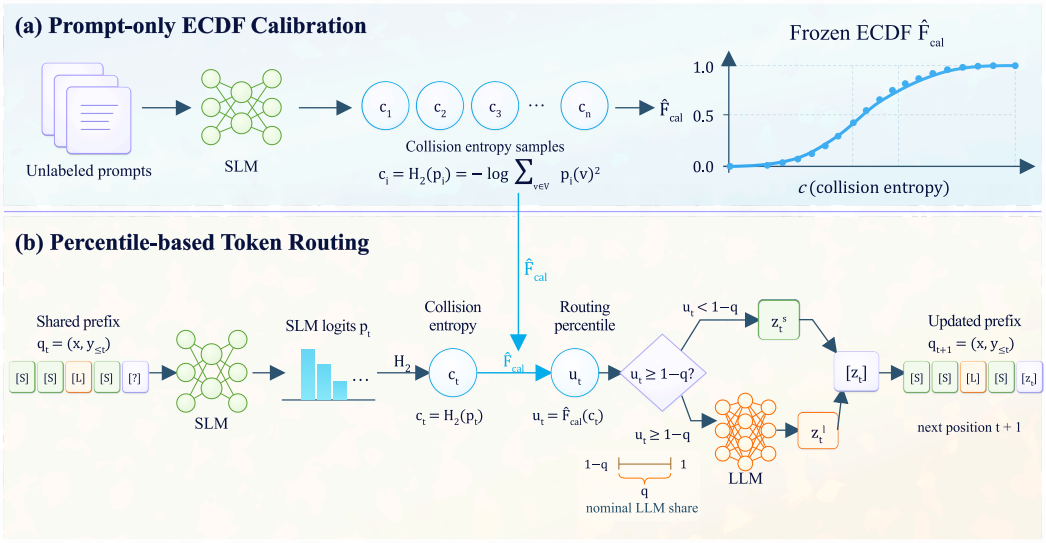


Figure 1: QUOTA decouples head-sensitive risk ranking from budget-interpretable token routing. (a) Prompt-only ECDF calibration maps collision-entropy scores from unlabeled SLM trajectories to a frozen percentile reference. (b) During inference, quota q selects the top-q risk tail; the LLM emits the next token for selected states and the SLM otherwise, after which routing is recomputed from the updated prefix.

3.2 PROBLEM FORMULATION

Let x denote the input prompt and $y_{(<t)}$ denote the shared output prefix generated up to position t. An SLM M_s and a stronger LLM M_l generate text according to the same task specification and share the same prefix history. QUOTA computes routing decisions from the SLM's next-token probability distribution over vocabulary V:

$$p_t(v) = p_s(v \mid x, y_{(<t)}), \qquad v \in V. \tag{1}$$

At position t, an action $a_t \in \{s, l\}$ selects which model emits the visible token z_t. The token is appended to the shared prefix to form $y_{(\leq t)}$, and the routing problem is re-solved at position $t+1$. Letting N_s and N_l denote the total number of tokens emitted by the SLM and LLM, respectively, the realized LLM token share is $r_l = \frac{N_l}{N_s+N_l}$. Rather than exposing a model-specific uncertainty threshold, QUOTA exposes a requested LLM token quota $q \in [0, 1]$ as the primary control parameter. The routing mechanism must satisfy two core desiderata: (i) it ranks generation states such that LLM capacity is concentrated on locally uncertain positions; (ii) it maps q to the ranking such that the parameter carries direct resource semantics across model pairs and workloads.

3.3 Head-Sensitive Risk Ranking and Quota Calibration

At each shared prefix, QUOTA computes collision entropy, the Rényi entropy of order 2, from the full SLM next-token distribution (Rényi, 1961):

$$c_t = H_2(p_t) = -\log \sum_{v \in V} p_t(v)^2. \tag{2}$$

Squaring the probabilities amplifies the contribution of high-probability leading candidates while diminishing the impact of diffuse low-probability tail tokens. As a result, c_t is low when a small number of plausible tokens dominate the distribution, and increases as competition among top candidates intensifies. Fragmentation of the low-probability tail has progressively less effect on H_2, as analyzed formally in Appendix A.3. This head-sensitive property makes collision entropy well suited for per-token routing decisions. The score is computed directly from the SLM's native logits and requires no auxiliary router, hidden-state probing, paired generation rollouts, or LLM queries.

Raw collision-entropy values lie on a numerical scale that varies with model size and workload. QUOTA calibrates them into model-relative percentiles using a small set of unlabeled calibration prompts D_{cal}. The SLM autoregressively processes each calibration prompt and collects collision-entropy samples $C = \{c_i\}_{i=1}^n$ across all visited token states. The ECDF is defined as

$$\hat{F}_{\text{cal}}(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[c_i \leq z], \tag{3}$$

and the calibrated percentile score is

$$u_t = \hat{F}_{\text{cal}}(c_t) \in [0, 1]. \tag{4}$$

Because $\hat{F}_{\text{cal}}$ is monotonic non-decreasing, u_t preserves the relative ordering of collision entropy scores while replacing model-specific numerical units with percentile coordinates. Calibration uses only unlabeled prompts and SLM-generated distributions: no reference answers, correctness labels, paired continuations, verifier calls, or routing labels are required. The frozen ECDF is built once per SLM configuration and reused across all online inference requests.

QUOTA expresses the user-specified LLM token quota $q \in [0, 1]$ directly on the calibrated percentile scale. For a requested quota q, the selected high-risk region is the upper-q tail of the reference ranking; equivalently, a token state is selected when

$$u_t \geq 1 - q. \tag{5}$$

For example, $q = 0.10$ selects states at or above the 90th percentile of the frozen calibration distribution. Because the cutoff operates on percentile rank, varying q expands or contracts the selected high-risk region without changing the collision-entropy ordering or retuning a model-specific raw threshold. Collision entropy therefore determines the priority of token states, the ECDF provides a model-relative ranking coordinate, and q directly specifies the requested allocation level. Section 3.4 defines the one-token action applied to selected states, while Section 4.3 evaluates the correspondence between requested and realized LLM usage across model pairs.

3.4 IMMEDIATE ONE-TOKEN DECISION

QUOTA aligns the action horizon with the scoring horizon: each selected token state triggers exactly one model decision. If $u_t \geq 1 - q$, the LLM receives the current shared prefix and emits the next token z_t; otherwise, the SLM emits z_t. The emitted token is appended to the shared prefix, and the entire routing decision is recomputed at position $t + 1$.

This one-token decision rule avoids the core limitation of step-level routing methods, which commit to an entire reasoning span based on a single local uncertainty estimate. An LLM token is often sufficient to redirect the shared prefix onto a correct reasoning path before the next routing decision, while the router immediately returns to SLM generation at routine low-uncertainty positions. This behavior is motivated by the same-prefix disagreement analysis in Appendix A.2, which quantifies how SLM–LLM output divergence is concentrated at individual token positions; related token-level collaboration methods likewise intervene selectively during decoding (Shen et al., 2024; Zheng et al., 2025; Fu et al., 2025; She et al., 2025b).

3.5 INFERENCE PROCEDURE

At deployment time, QUOTA repeats a score-map-route-update cycle at every token position. Once the ECDF is frozen for a given SLM, the user-specified quota q is the only deployment-time control parameter. Appendix C provides the complete online procedure.

4 EXPERIMENTS

Our evaluation tests four core hypotheses: (i) selective token routing improves reasoning quality under practical LLM compute allocations; (ii) the requested quota q maps reliably to realized LLM token usage across model pairs; (iii) performance gains stem from risk-aware placement of LLM capacity rather than total LLM usage alone; and (iv) head-sensitive ranking, ECDF calibration, and one-token decision jointly determine the final operating point.

4.1 EXPERIMENTAL SETUP

Models and inference infrastructure. Our primary end-to-end infrastructure uses Qwen3-0.6B as the SLM and Qwen3-32B as the LLM (Yang et al., 2025). Mechanism studies use the Qwen3-0.6B/8B pair, while cross-model budget transfer experiments additionally include DeepSeek-R1-Distill-Qwen-1.5B/32B (Guo et al., 2025). The default decoding configuration uses Qwen3's native reasoning mode with a thinking budget of 8192 tokens, temperature=0.7, and top-p=0.95. All experiments are implemented using the vLLM inference engine (Kwon et al., 2023) and run on four NVIDIA A100-40GB GPUs. Within each comparison, all methods share identical prompt template, decoding configuration, output length limit, and hardware environments.

Benchmarks and data partitions. Final evaluation spans three reasoning domains: mathematical reasoning on pooled AIME 2024 and AIME 2025, scientific question answering on GPQA-Diamond (Rein et al., 2024), and code generation on LiveCodeBench v6 (Jain et al., 2025). The ECDF is calibrated on a held-out partition of Bespoke-Stratos-17k using only unlabeled prompts and SLM-generated token distributions. We reserve a validation set consisting of AIME 2023, 69 held-out code problems from Bespoke-Stratos-17k, and 60 GPQA Extended questions (excluding GPQA-Diamond samples). Calibration, validation, and test partitions are strictly disjoint.

Baselines. We include standalone SLM and LLM endpoints as lower and upper bounds, plus Qwen3-8B as an intermediate single-model reference point. For collaborative inference baselines, we compare against SpecReason (Pan et al., 2025) and GlimpRouter (Zeng et al., 2026). QUOTA-MR is a matched-budget random control that preserves the same LLM token quota as QUOTA but randomizes the positions of LLM takeover. Component studies compare Shannon entropy H_1 and collision entropy H_2 under either raw numerical thresholds or ECDF percentiles.

Evaluation metrics. Accuracy is measured as pass@1 over four independent single-sample runs: each problem receives one sampled response per run, and benchmark-level success rates are averaged across runs. AIME and GPQA use answer-correctness grading; LiveCodeBench uses benchmark-test success evaluation. Latency is measured as end-to-end wall-clock time in seconds

per question, averaged over all requests and all four runs per benchmark. All "Average" columns report an equal-weight macro-average across benchmark-level mean scores, rather than a pooled request average. All reported accuracy and latency values are means over four complete runs.

For budget fidelity measurements, the realized LLM token share is computed as the aggregate token-level ratio across the full evaluation dataset:

$$r_l = \frac{\sum_i N_{l,i}}{\sum_i (N_{s,i} + N_{l,i})}. \tag{6}$$

Budget fidelity is evaluated across three model pairs and four requested quotas $q \in \{5\%, 10\%, 20\%, 30\%\}$, yielding 12 total operating points. Budget error is reported as mean absolute error across all operating points, in percentage points:

$$\mathrm{MAE}_{\mathrm{budget}} = \frac{1}{12} \sum_{j=1}^{12} |r_{l,j} - q_j|. \tag{7}$$

4.2 Comparison for Qwen3-0.6B/32B Model Pair

Table 1 presents the primary end-to-end comparison for the Qwen3-0.6B/32B model pair. QUOTA achieves a macro-average pass@1 of 48.55, the highest among all evaluated collaborative methods, with a mean latency of 167 seconds per question. Relative to SpecReason, QUOTA improves macro pass@1 by 2.90 percentage points while reducing mean latency by 17 seconds per question. QUOTA also outperforms both SpecReason and GlimpRouter on AIME, GPQA-Diamond, and LiveCodeBench v6.

Notably, QUOTA reaches performance parity with the standalone Qwen3-8B single-model baseline (48.55 vs. 48.68 pass@1) while reducing mean latency from 182 to 167 seconds per question. This demonstrates the core efficiency advantage of QUOTA: strong reasoning quality delivered via selective, fine-grained access to a larger model, rather than uniform full-model decoding.

Table 1: Main Qwen3-0.6B/32B results. Acc is mean pass@1 (%), Lat is mean end-to-end seconds/question, and Average is the equal-weight mean over pooled AIME24/25, GPQA-Diamond, and LiveCodeBench v6. All values are averages over four runs.

	AIME		LCBv6		GPQA		Average	
Method	Acc↑	Lat↓	Acc↑	Lat↓	Acc↑	Lat↓	Acc↑	Lat↓
Qwen3 0.6B	7.92	115	6.57	132	16.41	102	10.30	116
Qwen3 32B	53.75	232	47.57	259	59.85	198	53.72	230
Qwen3-8B	46.67	197	44.71	185	54.67	164	48.68	182
SpecReason	43.75	183	41.43	197	51.77	173	45.65	184
GlimpRouter	46.67	156	42.29	163	54.67	153	47.88	157
QUOTA-MR	30.83	217	31.71	231	24.37	187	28.97	212
QUOTA	47.50	163	42.71	181	55.43	158	48.55	167

Full quality–allocation curves across varying quota values are provided in Appendix B. As q changes, the token-level routing policy interpolates smoothly between lower-cost and higher-capacity regimes across mathematics, scientific question answering, and code generation.

4.3 Budget Fidelity across Model Pairs

Figure 2 evaluates whether the requested quota q translates predictably to realized LLM token share. Across Qwen3-0.6B/8B, Qwen3-0.6B/32B (Yang et al., 2025), and DeepSeek-R1-Distill-Qwen-1.5B/32B (Guo et al., 2025), realized LLM shares closely track the requested quotas $q \in \{5\%, 10\%, 20\%, 30\%\}$. The mean absolute budget error across all 12 model-pair/operating-point combinations is only 1.32 percentage points. This result confirms that the quota interface retains

consistent resource semantics across different SLM checkpoints, LLM scales, and model families, requiring only prompt-only ECDF calibration per SLM with no additional tuning.

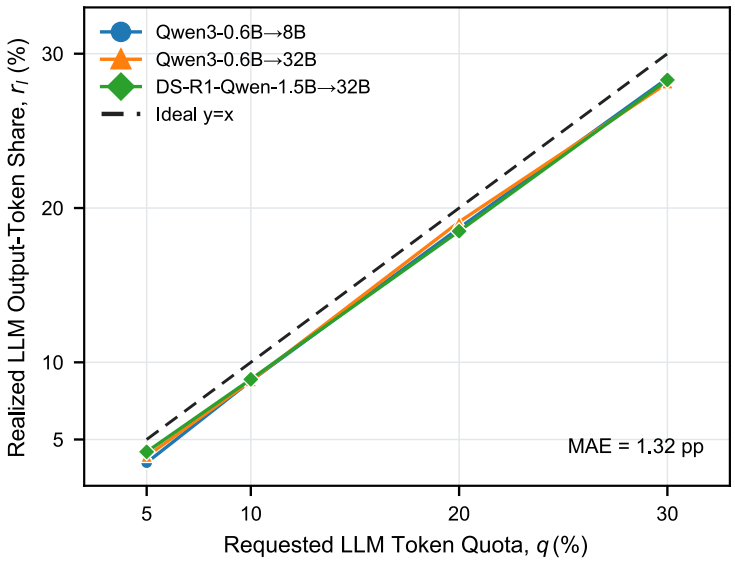


Figure 2: Nominal-to-realized LLM token share across three model pairs. The pooled MAE over 12 operating points is 1.32 percentage points; the dashed line is ideal allocation.

4.4 Risk-Aware Placement at Comparable LLM Allocation

The performance gains of selective routing depend critically on whether LLM capacity is deployed at informative positions. The QUOTA-MR matched-budget control isolates this effect by preserving the same total LLM budget quota but randomizing takeover locations. At an identical budget, QUOTA achieves 48.55 macro pass@1 compared to 28.97 for QUOTA-MR, a 19.58 percentage point gap. This large difference under matched compute provides direct causal evidence that QUOTA's improvements stem from risk-aware placement of LLM capacity, not simply from increased LLM usage overall.

Table 2 provides a complementary comparison on AIME25 with the Qwen3-0.6B/8B pair at similar LLM allocation levels. QUOTA reaches 42.50% accuracy while assigning 29% of output tokens to the LLM, compared to 40.83% at 30% LLM share for GlimpRouter and 37.50% at 33% LLM share for SpecReason. QUOTA thus converts a comparable or smaller LLM allocation into higher task accuracy, directly validating the benefits of token-level risk ranking under controlled budgets.

Table 2: AIME25 results at comparable LLM output-token share for Qwen3-0.6B/8B. Acc and Lat are averages over four runs; LLM Tok. is the aggregate output-token share.

Model Pair	Method	AIME25 Acc.	LLM Tok.↓	Lat↓
Qwen3 0.6B→8B	Pure SLM	6.67	0	115
	Pure LLM	43.33	100	197
	SpecReason	37.50	33	173
	GlimpRouter	40.83	30	162
	QUOTA	42.50	29	168

4.5 Ablation Study

We conduct factorial ablation studies to quantify the contribution of each core design component.

Uncertainty score and budget interface. Table 3 isolates the effects of ECDF calibration and head-sensitive scoring. ECDF calibration reduces budget MAE from 9 to 3 percentage points for Shannon entropy H_1, and from 7 to 1 for collision entropy H_2. The combination of H_2 scoring and ECDF calibration yields both the highest AIME25 accuracy (42.50%) and the lowest budget error (1 percentage point). Under the same ECDF interface, H_2-ECDF achieves 42.50% accuracy versus 37.50% for H_1-ECDF at similar realized LLM shares (29% vs. 27%), providing direct empirical support for the head-sensitive ranking principle underlying QUOTA.

Table 3: Shannon versus collision entropy with raw versus ECDF thresholds on AIME25 using Qwen3-0.6B/8B. Budget MAE is in percentage points.

Variant	AIME25 Acc.	LLM Tok. (%)	Budget MAE (pp)↓
H1-Raw	33.33	21	9
H1-ECDF	37.50	27	3
H2-Raw	34.17	23	7
H2-ECDF / QUOTA	42.50	29	1

Calibration efficiency. Table 4(a) evaluates performance as a function of calibration set size. With only 256 unlabeled prompts, QUOTA already achieves 42.50% accuracy and 1 percentage point budget MAE. Increasing to 512 prompts retains the 1-point MAE and further improves accuracy to 43.33%. This demonstrates that prompt-only calibration is highly sample-efficient, requiring minimal unlabeled data to achieve stable, accurate budget control.

Action granularity. Table 4(b) validates the immediate one-token takeover design. A one-token action span yields the highest accuracy (42.50%), lowest latency (168 seconds per question), and smallest LLM share (29%) among all tested spans. This result supports the design principle of aligning the action horizon exactly with the local risk scoring horizon.

Table 4: Calibration-size and takeover-span analysis on AIME25 with Qwen3-0.6B/8B. Acc and Lat are averages over four runs.

(a) Calibration Size

Calibration Prompts	AIME25 Acc.↑	Budget MAE (pp)↓
64	36.67	4
128	38.33	2
256	42.50	1
512	43.33	1

(b) Takeover Granularity

LLM Takeover Span	AIME25 Acc.↑	Lat↓	LLM Tok.↓
1 token	42.50	168	29
2 tokens	41.67	182	38
4 tokens	39.17	206	43

5 Conclusion

In this work, we present QUOTA, a training-free token-level routing framework that recasts small–large model collaborative inference as a principled resource allocation problem by decoupling where to invoke the LLM from how much capacity to allocate. Collision entropy provides head-sensitive state ranking, ECDF calibration enables interpretable quota semantics, and one-token takeover aligns action granularity with local risk. With Qwen3-0.6B/32B, QUOTA reaches 48.55 macro-average pass@1, 2.90 points above SpecReason with 17 seconds lower per-question latency. On AIME25 with Qwen3-0.6B/8B, it achieves 42.50% accuracy at 29% LLM token share; across three model pairs, budget MAE is 1.32 percentage points. Local uncertainty thus directs where strong-model compute is deployed, while a calibrated quota governs its magnitude, yielding a practical, training-free framework for efficient high-quality LLM inference.

AI USE STATEMENT

In this work, we used generative AI tools only to edit manuscript text for linguistic fluency, logical coherence, and readability. We did not use generative AI for any other task for which disclosure is recommended under the ICLR 2027 policy.

We did not use generative AI to collect, generate, clean, or reformat data; develop theoretical models or conceptual frameworks; formulate mathematical claims; contribute to or write proofs; propose or refine hypotheses; design or provide feedback on research methodology or experiments; implement methods; translate content; conduct qualitative or thematic data analysis; or interpret results. Any remaining tasks requiring disclosure were not applicable to this work.

All AI-assisted edits were reviewed and, where necessary, revised by the authors. The edited text was checked against the research design, implemented methods, experimental results, and intended claims to ensure accuracy and consistency and to prevent unsupported statements, misleading wording, or unattributed material. The authors take full responsibility for the final content of this work, including all text, claims, conclusions, and artifacts produced or revised with the aid of generative AI.

REFERENCES

Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav Gulavani, Alexey Tumanov, and Ramachandran Ramjee. Taming Throughput-Latency tradeoff in LLM inference with Sarathi-Serve. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pp. 117–134, Santa Clara, CA, July 2024. USENIX Association. ISBN 978-1-939133-40-3. URL https://www.usenix.org/conference/osdi24/presentation/agrawal.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5209–5235. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cai24b.html.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=cSimKw5p6R. Featured Certification.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks Lakshmanan, and Ahmed H Awadallah. Hybrid llm: Cost-efficient and quality-aware query routing. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 41348–41366, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/b47d93c99fa22ac0b377578af0a1f63a-Paper-Conference.pdf.

Dujian Ding, Ankur Mallick, Shaokun Zhang, Chi Wang, Daniel Madrigal, Mirian Del Carmen Hipolito Garcia, Menglin Xia, Laks V. S. Lakshmanan, Qingyun Wu, and Victor Rühle. BEST-route: Adaptive LLM routing with test-time optimal compute. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 13870–13884. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ding25d.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Tao Feng, Yanzhen Shen, and Jiaxuan You. Graphrouter: A graph-based router for llm selections. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 26186–26203, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/41b6674c28a9b93ec8d22a53ca25bc3b-Paper-Conference.pdf.

Tianyu Fu, Yi Ge, Yichen You, Enshu Liu, Zhihang Yuan, Guohao Dai, Shengen Yan, Huazhong Yang, and Yu Wang. R2r: Efficiently navigating divergent reasoning paths with small-large model token routing. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 124108–124145. Curran Associates, Inc., 2025. doi: 10.52202/085713-4135. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b39cef2ef90591cffdc9c674cd55bebe-Paper-Conference.pdf.

Yichao Fu, Peter Bailis, Ion Stoica, and Hao Zhang. Break the sequential dependency of LLM inference using lookahead decoding. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14060–14079. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/fu24a.html.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.366. URL https://aclanthology.org/2024.naacl-long.366/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Zhenyu He, Zexuan Zhong, Tianle Cai, Jason Lee, and Di He. REST: Retrieval-based speculative decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 1582–1595, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.88. URL https://aclanthology.org/2024.naacl-long.88/.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. Uncertainty in language models: Assessment through rank-calibration. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 284–312, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.18. URL https://aclanthology.org/2024.emnlp-main.18/.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3600006.3613165.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19274–19286. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/leviathan23a.html.

Xianzhi Li, Ethan Callanan, Abdellah Ghassel, and Xiaodan Zhu. Entropy-gated branching for efficient test-time reasoning. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5054–5069, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.235. URL https://aclanthology.org/2026.eacl-long.235/.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. EAGLE: Speculative sampling requires rethinking feature uncertainty. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28935–28948. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/li24bt.html.

Baohao Liao, Yuhui Xu, Hanze Dong, Junnan Li, Christof Monz, Silvio Savarese, Doyen Sahoo, and Caiming Xiong. Reward-guided speculative decoding for efficient LLM reasoning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37555–37572. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/liao25f.html.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601,

2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Putra Manggala, Atalanti A Mastakouri, Elke Kirschbaum, Shiva Kasiviswanathan, and Aaditya Ramdas. Qa-calibration of language model confidence scores. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 82631–82654, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/cd96cb9a239c37b39dbf34f3f5a4c56f-Paper-Conference.pdf.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, Chunan Shi, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, and Zhihao Jia. Specinfer: Accelerating large language model serving with tree-based speculative inference and verification. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, ASPLOS '24, pp. 932–949. ACM, April 2024. doi: 10.1145/3620666.3651335. URL http://dx.doi.org/10.1145/3620666.3651335.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Kadous, and Ion Stoica. Routellm: Learning to route llms from preference data. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 34433–34448, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5503a7c69d48a2f86fc00b3dc09de686-Paper-Conference.pdf.

Rui Pan, Yinwei Dai, Zhihao Zhang, Gabriele Oliaro, Zhihao Jia, and Ravi Netravali. Specreason: Fast and accurate inference-time compute via speculative reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 12730–12749. Curran Associates, Inc., 2025. doi: 10.52202/085713-0428. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/12c45a68e8433b21b91cd47731387fa4-Paper-Conference.pdf.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pp. 547–562. University of California Press, 1961.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 60808–60838, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/98711dea460bdefe0e651ca23ec98ba2-Paper-Conference.pdf.

Jianshu She, Zhuohao Li, Zhemin Huang, Qi Li, Peiran Xu, Haonan Li, and Qirong Ho. Hawkeye: Model collaboration for efficient reasoning. In *Second Conference on Language Modeling*, 2025a.

Jianshu She, Wenhao Zheng, Zhengzhong Liu, Hongyi Wang, Eric P. Xing, Huaxiu Yao, and Qirong Ho. Token level routing inference system for edge devices. In Pushkar Mishra, Smaranda Muresan, and Tao Yu (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 159–166, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-253-4. doi: 10.18653/v1/2025.acl-demo.16. URL https://aclanthology.org/2025.acl-demo.16/.

Zejiang Shen, Hunter Lang, Bailin Wang, Yoon Kim, and David Sontag. Learning to decode collaboratively with multiple language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12974–12990, Bangkok, Thailand, August

2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.701. URL https://aclanthology.org/2024.acl-long.701/.

Junhan Shi, Yijia Zhu, Zhenning Shi, Dan Zhao, Qing Li, and Yong Jiang. SpecCoT: Accelerating chain-of-thought reasoning through speculative exploration. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24405–24415, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1326. URL https://aclanthology.org/2025.findings-emnlp.1326/.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 10131–10165, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1b623663fd9b874366f3ce019fdfdd44-Paper-Conference.pdf.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. Calibrating large language models using their generations only. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15440–15459, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.824. URL https://aclanthology.org/2024.acl-long.824/.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 115452–115486. Curran Associates, Inc., 2025. doi: 10.52202/085713-3850. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a797c2d2e0c1fdabf4d1ab8cd0b465c6-Paper-Conference.pdf.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 55824–55845, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8c3caae2f725c8e2a55ecd600563d172-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Wenhao Zeng, Xuteng Zhang, Yuling Shi, Chao Hu, Yuting Chen, Beijun Shen, and Xiaodong Gu. GlimpRouter: Efficient collaborative inference by glimpsing one token of thoughts. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17850–17864, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.885. URL https://aclanthology.org/2026.findings-acl.885/.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. Calibrating the confidence of large language models by eliciting fidelity. In Yaser Al-

Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2959–2979, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.173. URL https://aclanthology.org/2024.emnlp-main.173/.

Xuechen Zhang, Zijian Huang, Ege Onur Taga, Carlee Joe-Wong, Samet Oymak, and Jiasi Chen. Efficient contextual llm cascades through budget-constrained policy learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 91691–91722. Curran Associates, Inc., 2024b. doi: 10.52202/079017-2910. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a6deba3b2408af45b3f9994c2152b862-Paper-Conference.pdf.

Wenhao Zheng, Yixiao Chen, Weitong Zhang, Souvik Kundu, Yun Li, Zhengzhong Liu, Eric P. Xing, Hongyi Wang, and Huaxiu Yao. CITER: Collaborative inference for efficient large language model decoding with token-level routing. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=nqX9UYW9Af.

Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. DistServe: Disaggregating prefill and decoding for goodput-optimized large language model serving. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pp. 193–210, Santa Clara, CA, July 2024. USENIX Association. ISBN 978-1-939133-40-3. URL https://www.usenix.org/conference/osdi24/presentation/zhong-yinmin.

A PRELIMINARY ANALYSIS

The design of QUOTA follows from three observations about small-large model collaboration. First, useful model differentiation is localized rather than uniform along a reasoning trajectory, which favors token-granular allocation. Second, an effective routing score should capture competition among plausible next tokens without being dominated by the cardinality of a diffuse vocabulary tail. Third, a deployment-facing threshold should retain consistent resource semantics when uncertainty scales vary across models and workloads. The following analyses establish these observations and connect them to the two core components of QUOTA: head-sensitive risk ranking and percentile-based budget calibration.

A.1 STUDY DESIGN

Unless otherwise stated, Figures A1–A3 use Qwen3-0.6B as the small language model (SLM) and Qwen3-8B as the large language model (LLM) (Yang et al., 2025). The analysis covers mathematical reasoning on AIME 2023, scientific question answering on GPQA Extended, and code generation on LiveCodeBench v4 under a common decoding protocol. For each local comparison, both models receive exactly the same generated prefix. This controlled same-prefix protocol isolates the decision faced by a token router: whether the SLM and LLM prefer the same next-token continuation from an identical reasoning state. All entropy statistics are computed from the complete SLM next-token distribution before sampling.

This design isolates three factors: the granularity at which model choice becomes relevant, the uncertainty statistic used to rank local states, and the parameterization that converts the statistic into an LLM allocation budget.

A.2 SPARSE BUT STRUCTURED TOKEN-LEVEL DISAGREEMENT

A central premise of fine-grained collaboration is that strong-model capacity need not be allocated uniformly throughout a reasoning trajectory. We evaluate this premise by measuring same-prefix top-1 disagreement between the SLM and LLM across relative generation positions.

Figure A1 shows that same-prefix disagreement is sparse across all three domains but varies substantially with the task and progress through the response. Mathematical reasoning, scientific question answering, and code generation exhibit distinct disagreement profiles rather than a common fixed region of model divergence. These profiles reveal a structured allocation opportunity: most positions exhibit local agreement, whereas a smaller subset contains divergent model preferences associated with reasoning-path changes in token-routing studies (Fu et al., 2025). Related work likewise identifies high-entropy minority tokens as consequential reasoning forks (Wang et al., 2025).

This concentration favors token-level routing over coarser allocation. Query-level escalation spends LLM capacity across entire responses, including positions with local agreement, while fixed step boundaries couple multiple token decisions into a single allocation. Token-level routing can instead concentrate LLM capacity at the localized positions where model choice can affect the continuation.

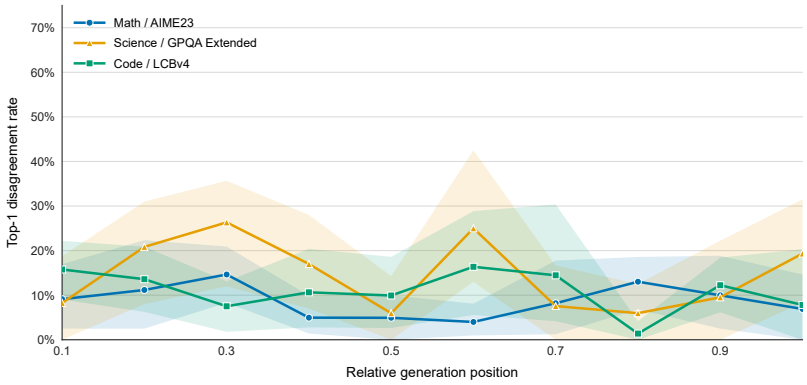


Figure A1: Sparse, structured same-prefix disagreement between Qwen3-0.6B and Qwen3-8B. Both models are evaluated from identical prefixes on AIME 2023, GPQA Extended, and LiveCodeBench v4. Disagreement occurs at a minority of positions and follows domain-dependent trajectories, motivating token-granular allocation of LLM capacity.

A.3 Shannon Entropy Overweights Tail Fragmentation

Effective token-level routing also requires a local uncertainty score aligned with the decision at each position. Predictive entropy and related measures are widely used to quantify uncertainty in language-model generations (Kuhn et al., 2023; Farquhar et al., 2024; Huang et al., 2024; Ulmer et al., 2024; Zhang et al., 2024a; Manggala et al., 2025; Geng et al., 2024; Wang et al., 2025). A common choice is Shannon entropy,

$$H_1(p) = -\sum_{v \in V} p(v) \log p(v), \tag{8}$$

which aggregates uncertainty over the entire vocabulary. For next-token routing, however, competition among the highest-probability candidates primarily determines the local decision. Shannon entropy can therefore become misaligned with the routing objective when a substantial vocabulary tail contains many individually negligible tokens.

To isolate this sensitivity, we fix the leading probabilities and total tail mass ε while distributing the tail mass uniformly over m low-probability tokens. The contribution of the tail to Shannon entropy is

$$H_{1,\text{tail}} = -m(\varepsilon/m)\log(\varepsilon/m) = -\varepsilon \log \varepsilon + \varepsilon \log m, \tag{9}$$

which increases logarithmically with the number of tail tokens. Thus, H_1 grows with tail support even though the distribution head, which governs competition among plausible next tokens, remains unchanged.

Collision entropy, the order-2 Rényi entropy,

$$H_2(p) = -\log \sum_{v \in V} p(v)^2, \tag{10}$$

responds differently. Under the same construction, the contribution of the tail to the collision probability is only

$$m(\varepsilon/m)^2 = \varepsilon^2/m. \tag{11}$$

As m grows, the contribution of the tail vanishes, leaving the collision probability and therefore H_2 primarily governed by concentration in the distribution head.

Figure A2 demonstrates this contrast: Shannon entropy continues to increase as tail support expands, whereas collision entropy remains comparatively stable once the head dominates. This behavior gives H_2 a head-sensitive inductive bias appropriate for one-token routing: the score emphasizes competition among plausible next tokens while suppressing variation caused by the size of a diffuse low-probability tail.

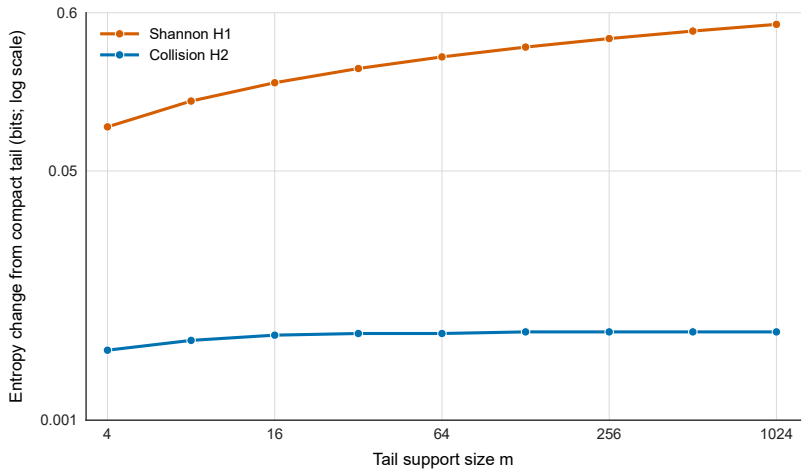


Figure A2: Collision entropy is less sensitive to vocabulary-tail fragmentation. The distribution head and total tail mass remain fixed while the tail support increases. Shannon entropy rises with tail support, whereas collision entropy changes little because concentration in the distribution head dominates the score.

A.4 Raw Uncertainty Has Model-Dependent Budget Semantics

A suitable ranking statistic determines where LLM intervention should be prioritized, but a deployment-facing router must also specify how much LLM capacity to allocate.

SLM uncertainty distributions vary across model checkpoints, task domains, prompt formats, sequence positions, and decoding configurations. Recent work on confidence estimation and calibration likewise shows that uncertainty scores require careful interpretation across tasks and models (Huang et al., 2024; Ulmer et al., 2024; Zhang et al., 2024a; Manggala et al., 2025; Geng et al., 2024). Consequently, a fixed numerical entropy threshold can route different fractions of tokens under different operating conditions. The raw threshold therefore conflates two quantities: the ordering of local uncertainty and the scale of the score distribution.

This distinction is operationally important for collaborative inference. The routing interface should allow an operator to request a target LLM share in a coordinate with a consistent interpretation, avoiding repeated adjustment of a model-specific entropy cutoff for each deployment setting. This requirement motivates expressing uncertainty in distribution-relative percentile coordinates.

Given the frozen calibration distribution of SLM uncertainty scores, an empirical cumulative distribution function (ECDF) maps a raw score c_t to

$$u_t = \hat{F}_{\text{cal}}(c_t) \in [0, 1]. \tag{12}$$

The value u_t is the empirical percentile of the current state relative to the calibration distribution. Applying the routing rule $u_t \geq 1 - q$ converts a model-specific numerical score into an interpretable tail allocation and targets approximately the upper-q fraction of reference uncertainty states for LLM intervention.

This transformation separates the two roles: the uncertainty statistic ranks local states, whereas ECDF calibration provides the coordinate in which the routing budget is expressed.

A.5 Head-Sensitive Ranking Meets Budget Calibration

Figure A3 isolates the two design choices through a factorial comparison of the uncertainty statistic (Shannon entropy H_1 or collision entropy H_2) and threshold parameterization (raw score or ECDF percentile).

The four variants answer complementary questions. Comparing H_1 with H_2 tests the ranking statistic, while comparing raw with ECDF parameterization tests how the routing threshold translates into realized LLM usage. Across the evaluated operating points, H_2-ECDF follows the ideal nominal-to-realized budget line most closely, demonstrating the joint value of head-sensitive ranking and percentile calibration for accurate budget control.

This result validates the central decomposition of QUOTA. Collision entropy supplies a head-sensitive ordering of token states, while ECDF calibration converts that ordering into a budget-facing percentile coordinate. Because the ECDF transformation is monotone, it preserves the ordering induced by the underlying score and expresses the operating point in terms of relative tail mass. The two components therefore solve distinct requirements within a unified routing rule.

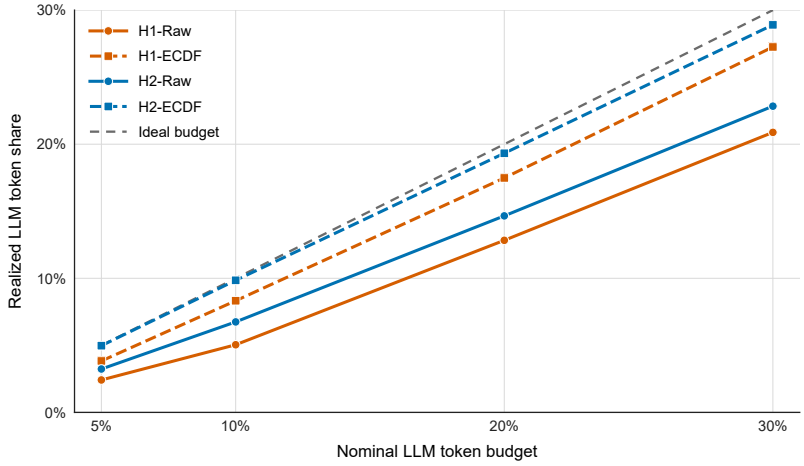


Figure A3: Factorial comparison of score design and budget parameterization. Realized LLM output-token share is plotted against nominal budget for H_1-Raw, H_1-ECDF, H_2-Raw, and H_2-ECDF on Qwen3-0.6B/8B. The dashed diagonal represents ideal budget fidelity; H_2-ECDF tracks this target most closely across the evaluated operating points.

A.6 From Analysis to QUOTA

Together, these analyses determine the design of QUOTA. Sparse, domain-dependent disagreement supports token-level model selection. The tail-fragmentation analysis identifies collision entropy as a head-sensitive local risk score, while the model-dependent scale of raw uncertainty motivates ECDF percentile calibration as the budget interface. Their combination allows the router to identify high-risk token states from the SLM distribution and express LLM intervention through an explicit nominal token budget, without training an auxiliary router or constructing token-level routing labels.

The resulting principle is direct: **collision entropy determines where LLM capacity is most valuable, while ECDF calibration determines how much LLM capacity is allocated.** Section 3 implements this principle through the immediate one-token routing policy of QUOTA.

B Quality-Allocation Curves

Figure B1 plots pass@1 against the requested LLM token quota q for Qwen3-0.6B/32B on AIME, GPQA-Diamond, and LiveCodeBench v6. Changing the requested quota moves the operating point

smoothly between lower-cost and higher-capacity regimes without retraining or changing the token-level routing rule.

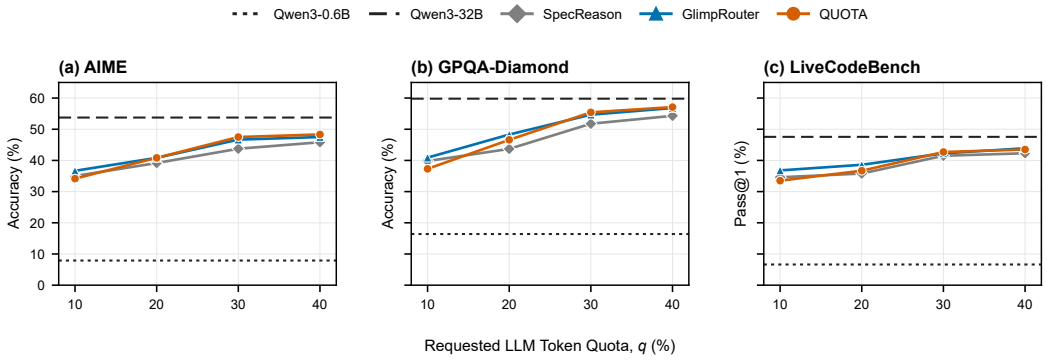


Figure B1: Quality-allocation curves for Qwen3-0.6B/32B. Pass@1 is plotted against the requested LLM token quota q on AIME, GPQA-Diamond, and LiveCodeBench v6; horizontal lines mark the standalone endpoints.

C Online Inference Pseudocode

Algorithm C1 specifies online QUOTA inference with the frozen prompt-only ECDF.

Algorithm C1 QUOTA Online Token Routing

Input: prompt x; SLM M_s; LLM M_l; frozen ECDF $\hat{F}_{\text{cal}}$; requested LLM token quota $q \in [0, 1]$; stopping rule.
Initialize: generated sequence $y \leftarrow \varnothing$; $N_s \leftarrow 0$; $N_l \leftarrow 0$.

1. While the stopping rule is not satisfied:
2. Evaluate the SLM next-token distribution

$$p_t(v) = p_s(v \mid x, y), \qquad v \in V. \tag{13}$$

3. Compute the head-sensitive risk score

$$c_t = -\log \sum_{v \in V} p_t(v)^2. \tag{14}$$

4. Map the score to its calibrated percentile

$$u_t = \hat{F}_{\text{cal}}(c_t). \tag{15}$$

5. If $u_t \geq 1 - q$, let M_l emit one token z_t and set $N_l \leftarrow N_l + 1$.
6. Otherwise, let M_s emit one token z_t using the current SLM distribution and set $N_s \leftarrow N_s + 1$.
7. Append z_t to the shared sequence: $y \leftarrow y \oplus z_t$.
8. End while.

Output: generated response y and realized LLM token share

$$r_l = \frac{N_l}{N_s + N_l}. \tag{16}$$