CRISP: CONFORMAL ROUTING AND INTERRUPTION WITH STATISTICAL PROTECTION FOR LLM CASCADES

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable deployment of large language model (LLM) cascades requires deciding both which model to query and whether its realized output should be returned. We introduce **CRISP** (**C**onformal **R**outing and **I**nterruption with **S**tatistical **P**rotection), which calibrates prompt- and output-side scores against labeled model failures. Under exchangeability, CRISP gives finite-sample, model- and task-conditional bounds on two events: an actually failing model entering the eligible set and an incorrect output passing the stopping gate. These are fixed-model failure-pass guarantees, not a bound of the same level on selective error or on the adaptively selected portfolio. Across seven open-weight models and four reasoning benchmarks, CRISP attains 81.5% accuracy and 99.5% coverage at a 0.0794 relative token-cost proxy on a fixed 200-example test split. Together, the formulation and evaluation show how failure-conditional calibration makes both routing and stopping decisions explicit and auditable in heterogeneous LLM cascades.

1 INTRODUCTION

LLMs are increasingly served through heterogeneous portfolios rather than as a single fixed model. Recent systems range from difficulty-aware weak/strong dispatch to preference-trained and dynamic multi-model routing (Ding et al., 2024; Ong et al., 2025; Wang et al., 2025b). Routing and cascading can reduce cost by sending easy prompts to smaller models and escalating difficult cases (Chen et al., 2024; Dekoninck et al., 2025). Reliability depends on whether a model suits the prompt and whether its realized output should be returned (Figure 1).

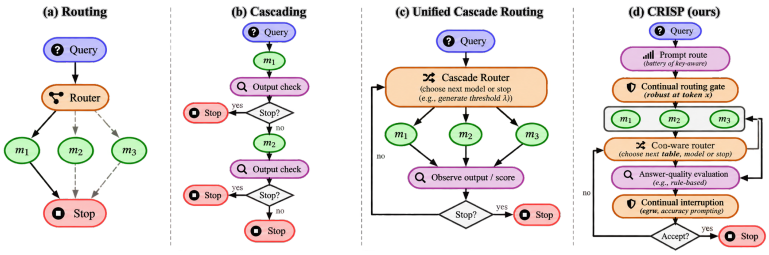


Figure 1: Model-selection strategies. CRISP applies separate conformal gates before selecting among candidate models and after observing each generated answer.

Benchmarking efforts now compare routers across increasingly large task and model pools (Hu et al., 2024; Huang et al., 2025; Li et al., 2026). They confirm useful model complementarity, but also show that performance depends strongly on the candidate pool, evaluator, and cost model. Current systems commonly use log-probability, self-consistency, or learned utility thresholds. Such scores can be predictive yet have different scales and failure semantics across models and tasks. A poor routing decision exposes a prompt to an unsuitable model; a poor stopping decision turns an incorrect candidate into the final answer. Cost–utility formulations clarify the optimization problem (Dekoninck et al., 2025), but do not control how often actual failures pass either gate.

Conformal methods offer finite-sample calibration under exchangeability and have been adapted to language-model prediction sets, black-box API access, and factuality-preserving backoff (Quach et al., 2024; Su et al., 2024; Mohri and Hashimoto, 2024). These methods primarily calibrate the

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.

uncertainty or content of one generator. A cascade additionally needs calibrated actions before generation and after observing a candidate.

We introduce **CRISP**, which calibrates both decisions against labeled failures. For every model–task pair, a prompt is eligible only when its risk score is unusually favorable relative to prompts that model answered incorrectly. After generation, the output is accepted only when its uncertainty score is unusually favorable relative to incorrect calibration outputs. Under exchangeability, the resulting lower-tail ranks control false eligibility and false acceptance for each fixed model. A union-bound extension is mathematically valid when error budgets are allocated across models, but its calibration requirements exceed the data sizes evaluated here. The guarantees deliberately concern fixed-model control events, not the error rate conditional on acceptance or the reported adaptive portfolio as a whole.

Our contributions are: (1) a failure-conditional conformal formulation for prompt routing and output interruption, with finite-sample fixed-model guarantees and an explicit analysis of rank resolution; (2) an entropy-focused NLL score instantiation; and (3) a seven-model, four-task case study with component ablations, a larger fixed split, prompt perturbations, and an out-of-distribution (OOD) stress test. On the primary split, CRISP achieves 81.5% accuracy and 99.5% coverage at 0.0794 relative token-cost proxy. We do not claim that the present experiments demonstrate a usable portfolio-wide error budget or practical latency savings.

2 Background and Related Work

Efficient Multi-LLM Inference. Modern LLM serving increasingly relies on *model portfolios* rather than a single monolithic model, trading quality against latency and monetary cost (Chen et al., 2024; Hu et al., 2024). Difficulty-aware and preference-trained routers learn one-shot model selection, while dynamic routers adapt assignments as workloads or model pools evolve (Ding et al., 2024; Ong et al., 2025; Wang et al., 2025b). The two dominant paradigms are *routing*, which selects one model from prompt features, and *cascading*, which queries models until a stop rule fires; unified utility-based views combine both (Dekoninck et al., 2025). Recent benchmarks broaden evaluation across tasks, model pools, and performance–cost objectives, while also documenting substantial headroom over current routers (Huang et al., 2025; Li et al., 2026). These methods nevertheless depend on quality estimates that are not statistically calibrated and can become brittle under distribution shift. Most closely related, CP-Router uses conformal prediction to route between an LLM and a large reasoning model based on prompt-side uncertainty (Su et al., 2026); CRISP differs by conformalizing the full cascade control loop, combining prompt-level eligibility with a post-generation interruption gate that verifies the realized output before accepting or escalating.

Reliability via Conformal Prediction. Reliable cascades require stop signals that are comparable across tasks and model families. Common heuristics—including raw log-probabilities, self-consistency, and learned verifiers—are often difficult to calibrate or interpret uniformly (Quach et al., 2024). Conformal prediction offers a distribution-free alternative with finite-sample guarantees under exchangeability (Barber et al., 2023). Recent language-model applications calibrate generation sets, black-box API prediction sets, and factuality-preserving backoff (Quach et al., 2024; Su et al., 2024; Mohri and Hashimoto, 2024). CRISP instead applies conformal calibration to the *control logic* of the system itself: routing eligibility and stop/escalation decisions.

Answer Scoring and Entropy-Gated Risk. The interruption rule also needs a scalar score that is informative about generation failure. Sequence-average negative log-likelihood (NLL) often washes out decision-critical uncertainty because long responses contain many low-entropy tokens. Recent analyses of *forking tokens* suggest that risk is concentrated in a small set of high-uncertainty generation steps (Wang et al., 2025a). Motivated by this, CRISP uses entropy-gated NLL, which scores only high-entropy tokens. The present experiments vary the retained entropy fraction but do not compare against ordinary sequence-average NLL ($\rho = 1$), entropy alone, or actual-failure AUROC; consequently they establish sensitivity, not superiority of the score.

Scope of Statistical Protection. CRISP calibrates two failure-pass events: whether a model that is wrong on the current prompt enters the eligible set, and whether an incorrect generated candidate passes the interruption gate. These guarantees do not directly bound the error rate among accepted

answers, which also depends on model failure prevalence and the gate's power on correct outputs. The current artifact reports accuracy, coverage, and a token-cost proxy, but lacks the held-out model–task failure table needed to audit the fixed-model pass rates empirically; the theoretical and empirical claims are therefore kept separate.

3 METHODOLOGY

3.1 PROBLEM SETUP

Let (X, Y, T) denote a random prompt–label–task triple and (x, y, t) a realization. We consider a portfolio $\mathcal{M} = \{M_1, \ldots, M_K\}$ of K models with increasing serving costs; $k \in \{1, \ldots, K\}$ indexes a model. Under a fixed decoding procedure, model M_k produces $\hat{y}_k(x)$, and

$$\ell_k(x, y) = \mathbb{1}[\hat{y}_k(x) \text{ fails the task-specific evaluator against } y] \tag{1}$$

denotes binary failure. We write $L_k = \ell_k(X, Y)$ and $\hat{Y}_k = \hat{y}_k(X)$ for the corresponding test-example random variables. Unless stated otherwise, probabilities are marginal over the exchangeable calibration/test sample and any fixed decoding randomization. At test time the controller chooses a model, observes its output, and either accepts, escalates to a more expensive eligible model, or abstains. CRISP calibrates these routing and stopping decisions; it does not claim assumption-free semantic correctness of arbitrary generated text.

We use a labeled calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i, t_i)\}_{i=1}^{n}$ and a disjoint test set. Task type t is supplied by each benchmark; in deployment, a task classifier must be fixed before calibration. Each model uses two smaller-is-better scalar statistics:

$$r(x) \in \mathbb{R} \qquad \text{(prompt-side risk proxy)}, \tag{2}$$

$$s_k(x, \hat{y}_k(x)) \in \mathbb{R} \qquad \text{(output-side uncertainty proxy)}. \tag{3}$$

Within task t, we standardize them using calibration means and standard deviations, with a fixed numerical stabilizer $\epsilon > 0$,

$$\tilde{r}(x; t) = \frac{r(x) - \mu_r(t)}{\max\{\sigma_r(t), \epsilon\}}, \tag{4}$$

$$\tilde{s}_k(x, \hat{y}; t) = \frac{s_k(x, \hat{y}) - \mu_{s,k}(t)}{\max\{\sigma_{s,k}(t), \epsilon\}}. \tag{5}$$

Because these are monotone transformations within each model–task group, they do not change the conformal ranks below.

3.2 QUALITY STATISTICS

Entropy-gated NLL. For a token sequence $y = (y_1, \ldots, y_L)$, let $\mathcal{I} \subseteq \{1, \ldots, L\}$ be the valid scored positions and $\mathcal{V}$ the model vocabulary. Let $p_j(v)$ be the next-token probability of token $v \in \mathcal{V}$ at position $j \in \mathcal{I}$, with entropy $H_j = -\sum_{v \in \mathcal{V}} p_j(v) \log p_j(v)$ and realized-token NLL $n_j = -\log p_j(y_j)$. For a global fraction $\rho \in (0, 1]$, let $\mathcal{S}_\rho \subseteq \mathcal{I}$ contain the $\lceil \rho |\mathcal{I}| \rceil$ positions with largest entropy. We define

$$\text{EG-NLL}_\rho(y) = \frac{1}{|\mathcal{S}_\rho|} \sum_{j \in \mathcal{S}_\rho} n_j. \tag{6}$$

The main experiments use $\rho = 0.2$, selected in the separate validation study in Appendix B.7. Calibration is applied after fixing ρ, so validity does not require EG-NLL to be a perfect correctness estimator.

Prompt statistic. Using a lightweight probe model P on prompt tokens $x_{1:L}$, we set

$$\begin{aligned} r(x) = \text{EG-NLL}_{0.2}(x_{2:L}) + c_{\text{in}}|x| \\ + c_{\text{dig}} \log(1 + \#\text{digits}(x)) + \delta_{\text{code}} \mathbb{1}[\text{code_flag}(x)], \end{aligned} \tag{7}$$

where $|x|$ is input length in thousands of tokens, $\#\text{digits}(x)$ counts digit characters, and $\text{code_flag}(x) \in \{0, 1\}$ indicates a code-like prompt. The fixed coefficients are weak priors rather than decision thresholds and are discussed in Section 4.1.

Output statistic. For greedy integer or code outputs, we use

$$s_k(x, \hat{y}) = \text{EG-NLL}_{0.2}(\hat{y}) + \text{penalty}(x, \hat{y}), \tag{8}$$

with a 0.2 penalty when no integer can be parsed and a 0.5 penalty when a required fenced code block is absent. For multiple-choice tasks, s_k is the negative top-1 versus top-2 option-probability margin. The output score is obtained from the same generation pass used to produce the candidate.

Low NLL can accompany a confidently wrong answer, so CRISP calibrates scores against observed model failures rather than treating the score itself as a correctness probability. The present evaluation does not report the previously proposed resampling-instability correlations or AUROC diagnostics.

3.3 Failure-conditional conformal calibration

For each model–task pair, define the calibration failure set

$$\mathcal{F}_{k,t} = \{i : t_i = t,\ \ell_k(x_i, y_i) = 1\}, \qquad m_{k,t} = |\mathcal{F}_{k,t}|. \tag{9}$$

For a smaller-is-better test value v and failure-reference values $v_{1:m}$, define the lower-tail conformal rank

$$\widehat{p}(v; v_{1:m}) = \frac{1 + \sum_{j=1}^{m} \mathbb{1}[v_j \le v]}{m+1}. \tag{10}$$

Using $\le$ assigns the upper rank within ties and is conservative. If $m = 0$, the rank equals one, so protected mode does not infer safety from an empty failure sample.

The routing and stopping ranks are

$$p^{\text{route}}_{k,t}(x) = \widehat{p}\big(\tilde{r}(x;t); \{\tilde{r}(x_i;t)\}_{i\in\mathcal{F}_{k,t}}\big), \tag{11}$$

$$p^{\text{stop}}_{k,t}(x, \hat{y}) = \widehat{p}\big(\tilde{s}_k(x, \hat{y}; t); \{\tilde{s}_k(x_i, \hat{y}_k(x_i); t)\}_{i\in\mathcal{F}_{k,t}}\big). \tag{12}$$

Small ranks mean that a prompt or output looks unusually favorable relative to examples on which that model was wrong. We assume that calibration and test examples, including decoding randomness, are exchangeable within each task group. With model-wise levels $\alpha_{\text{route},k}$ and $\alpha_{\text{stop},k}$, exchangeability gives

$$\mathbb{P}\big(p^{\text{route}}_{k,t}(X) \le \alpha_{\text{route},k} \mid \ell_k(X,Y) = 1, T = t\big) \le \alpha_{\text{route},k}, \tag{13}$$

$$\mathbb{P}\Big(p^{\text{stop}}_{k,t}(X, \hat{Y}_k) \le \alpha_{\text{stop},k} \mid \ell_k(X,Y) = 1, T = t\Big) \le \alpha_{\text{stop},k}. \tag{14}$$

Appendix A proves these statements. They control the fraction of actual failures that pass a gate, not the selective error $\mathbb{P}(\ell = 1 \mid \text{accept})$. The experiments use common per-model values $\alpha_{\text{route},k} = 0.05$ and $\alpha_{\text{stop},k} = 0.15$; smaller values are stricter.

Define the output accept event

$$A_k(x;t) = \mathbb{1}\Big[p^{\text{stop}}_{k,t}(x, \hat{y}_k(x)) \le \alpha_{\text{stop},k}\Big]. \tag{15}$$

3.4 Routing and interruption

For a test prompt with task type t, the protected eligible set is

$$\mathcal{E}(x,t) = \{k : p^{\text{route}}_{k,t}(x) \le \alpha_{\text{route},k}\}. \tag{16}$$

If $\mathcal{E}(x,t) = \emptyset$, protected-mode CRISP abstains. A deployment may instead query the largest model as an explicitly unprotected fallback, but such a return is outside the guarantees.

After querying eligible model M_k, CRISP returns its output iff

$$p^{\text{stop}}_{k,t}(x, \hat{y}_k(x)) \le \alpha_{\text{stop},k}. \tag{17}$$

Otherwise it escalates to the next model in the planned eligible suffix; if no candidate passes, it abstains. This realized-output check distinguishes CRISP from routing-only conformal methods.

3.5 EXPECTED-COST ROUTING

Let the $q = |\mathcal{E}(x,t)|$ eligible models be ordered by cost as $k_1, \ldots, k_q$, and let $n_t = |\{i : t_i = t\}|$ be the number of task-t calibration examples. Let $p_{k,t}^{\mathrm{stop},-i}$ denote the stopping rank computed after excluding example i from its failure-reference set. CRISP estimates the marginal acceptance probability

$$\widehat{p}_{k,t} = \frac{1}{n_t} \sum_{i:t_i=t} \mathbb{1}\left[p_{k,t}^{\mathrm{stop},-i}(x_i, \hat{y}_k(x_i)) \le \alpha_{\mathrm{stop},k}\right]. \tag{18}$$

Estimates may additionally be conditioned on a prespecified calibration-quantile bin $B(x)$ of $r(x)$; denote the resulting estimate by $\widehat{p}_{k,t,B(x)}$. Let $\widehat{c}_k(x)$ denote the estimated normalized cost of querying model M_k on prompt x.

For a possible starting position j, the estimated suffix cost obeys

$$\widehat{\mathrm{EC}}_j(x,t) = \widehat{c}_{k_j}(x) + \left(1 - \widehat{p}_{k_j,t,B(x)}\right) \widehat{\mathrm{EC}}_{j+1}(x,t), \tag{19}$$

with $\widehat{\mathrm{EC}}_q = \widehat{c}_{k_q}(x)$. CRISP starts at

$$j^\star = \arg\min_{1 \le j \le q} \widehat{\mathrm{EC}}_j(x,t), \qquad k^\star = k_{j^\star}, \tag{20}$$

and, if necessary, queries $k_{j^\star+1}, \ldots, k_q$ in order. Because acceptance events after earlier rejection can be correlated, Eq. (19) is a plug-in heuristic based on marginal acceptance probabilities, not an unbiased conditional expected-cost recursion unless an additional conditional model or independence assumption holds. This ordering affects efficiency only; the gate guarantees remain rank-based.

Reported cost proxy. For a queried set $Q(x)$, the implementation records a model-weighted token proxy of the form

$$C_{\mathrm{proxy}}(x) = \frac{1}{C_{\mathrm{ref}}} \sum_{k \in Q(x)} \left(w_k^{\mathrm{in}} N_k^{\mathrm{in}}(x) + w_k^{\mathrm{out}} N_k^{\mathrm{out}}(x)\right), \tag{21}$$

where C_{ref} is the implementation's normalization for one largest-model pass. The reported experimental record does not specify C_{ref}, the numerical weights, the probe identifier, or the per-query ledger, and it excludes complete latency/energy accounting for probing, prefill, routing, and score extraction. We therefore treat cost as a descriptive implementation-specific proxy rather than total compute or deployment latency.

3.6 GUARANTEE INTERPRETATION AND FINITE-SAMPLE RESOLUTION

From failure-pass control to selective error. The stopping guarantee has a precise but intentionally conditional meaning. Write $A_k = A_k(X;T)$ for the random acceptance indicator, and let $\pi_{k,t} = \mathbb{P}(L_k = 1 \mid T = t)$ be model k's base failure rate and $a_{k,t} = \mathbb{P}(A_k = 1 \mid T = t)$ its marginal acceptance probability. Whenever $a_{k,t} > 0$, Bayes' rule and Eq. (14) give

$$\mathbb{P}(L_k = 1 \mid A_k = 1, T = t) = \frac{\mathbb{P}(A_k = 1 \mid L_k = 1, T = t)\pi_{k,t}}{a_{k,t}} \le \frac{\alpha_{\mathrm{stop},k}\pi_{k,t}}{a_{k,t}}. \tag{22}$$

Thus, a small failure-pass level is useful but is not by itself a bound of the same numerical size on the error among accepted answers. Selective error also depends on model quality and on how often correct outputs pass. This distinction motivates reporting end-to-end accuracy and coverage alongside the calibrated event rates.

Discrete resolution. With $m_{k,t}$ calibration failures, the attainable ranks lie on the grid $\{1/(m_{k,t}+1), \ldots, 1\}$. For either gate, let α_k denote its model-wise routing or stopping level. A protected gate cannot pass any candidate unless

$$m_{k,t} \ge \left\lceil \frac{1}{\alpha_k} \right\rceil - 1. \tag{23}$$

At the main levels, this necessary condition is at least 19 failures for $\alpha_{\mathrm{route}} = 0.05$ and 6 failures for $\alpha_{\mathrm{stop}} = 0.15$. If a model–task group is smaller, CRISP is conservative: the corresponding gate

never passes in protected mode. In particular, a balanced total calibration size of 40 provides only 10 examples per task and must yield zero protected routing coverage; no such empirical result is retained. Pooling across nonexchangeable tasks merely to obtain finer ranks would invalidate the task-conditional argument, so larger calibration sets or a prespecified coarser grouping are preferable.

Fixed-model versus adaptive control. Equations (13)–(14) hold for each fixed model. Since CRISP adaptively chooses among models, a deployment that requires a single familywise guarantee can allocate global budgets across the portfolio. For $K = 7$ and uniform routing allocation $0.05/7$, however, Eq. (23) requires at least 139 failures per model–task pair; this is impossible with the 50 or 100 calibration examples per task used here. Appendix A.5 proves the union-bound statement, but the experiments use common per-model levels and do not demonstrate a nontrivial portfolio-wide guarantee.

3.7 Overall procedure

Offline calibration generates every model output, identifies failures, stores model–task failure-reference scores, and estimates leave-one-out acceptance probabilities. Online inference constructs $\mathcal{E}(x, t)$, selects the expected-cost-minimizing suffix, and accepts only an output that passes its stopping rank. Appendix C gives pseudocode. Appendix A proves fixed-model false-eligibility and false-acceptance control, a distribution-shift degradation bound, and a familywise adaptive statement when model-wise budgets sum to global budgets.

4 Results

We evaluate CRISP on seven open-weight LLMs and four benchmarks: HellaSwag Zellers et al. (2019), WinoGrande XL Sakaguchi et al. (2021), ARC-Challenge Clark et al. (2018), and GSM8K Cobbe et al. (2021). Comparisons include three routing/cascading baselines and component ablations. Experiments ran on three NVIDIA RTX A6000 GPUs (48 GB each).

4.1 Experimental Setup

Model Portfolio. We construct a seven-model cascade spanning three empirical cost–capability tiers. The tiers describe roles in this serving portfolio rather than disjoint parameter-count intervals:

- **Tier 1 (entry):** Qwen2.5-3B Qwen Team (2024), Phi-3.5-mini-instruct Microsoft (2024), Llama-3.1-8B Grattafiori et al. (2024).
- **Tier 2 (intermediate):** Qwen2.5-7B Qwen Team (2024), Qwen3-32B Yang et al. (2025), Dolphin-2.9.1-Yi-1.5-34B Hartford et al. (2024); Young et al. (2024).
- **Tier 3 (fallback):** Llama-3.3-70B-Instruct Meta (2024).

Datasets, tasks, and statistics. We stratify 400 examples (100 per dataset) into $N_{\text{cal}} = 200$ calibration and $N_{\text{test}} = 200$ evaluation cases spanning commonsense, reference resolution, science, and arithmetic. This is one fixed split, not a repeated-seed study; Wilson 95% intervals quantify test-count uncertainty where applicable, and no significance claim is made without paired outcomes. Appendix B.6 reports a larger disjoint split; Appendix B.4 audits which protected levels are even attainable at these calibration sizes.

Metrics. Accuracy counts abstentions as incorrect, coverage is the non-abstention fraction, and relative cost is the implementation's token proxy in Eq. (21). The proxy is useful for within-implementation accounting but is not a measurement of latency, energy, or total serving compute because the available artifact does not identify the probe, numeric token weights, or a complete per-query ledger.

Prompt-score coefficients. The terms added to $\text{EG-NLL}_{0.2}$ in $r(x)$ are weak score-design priors, not gate thresholds. We use c_{in}=0.0015 for input length, consistent with token-dependent serving cost (Dekoninck et al., 2025; OpenAI, 2026); c_{dig}=0.04 on the logarithm of the digit count as a diminishing numerical-difficulty prior (Gao et al., 2023; Mirzadeh et al., 2024); and δ_{code}=0.2 for prompts requiring functional code evaluation (Chen et al., 2021). Task-wise standardization and

conformal ranking determine the decisions; the coefficients are shared across datasets and are not tuned per task.

Component baselines. RoutingOnly uses CRISP's eligibility and cost policy without the output stopping gate; CascadingOnly changes both eligibility/order and the stopping rule by using a fixed sequence and fixed thresholds; CheapestFirst always queries the least expensive model; and RandomSingle selects one model at random. All use the same test split and proxy normalization. Consequently, only RoutingOnly is a near-matched comparison for the incremental stopping gate.

4.2 Main Results

On the fixed primary split (N_{test}=200), CRISP returns 163 correct answers, abstains once, and records 0.0794 mean proxy cost (Table 1). Wilson intervals reflect test-count uncertainty, not split-to-split or decoding variability. The larger disjoint split records 86.0% accuracy and 99.5% coverage (Appendix B.6). These are descriptive system outcomes: the reported logs omit the all-model, model–task failure events needed to audit the theorem empirically.

Table 1: Aggregate CRISP performance on the fixed primary split ($N_{\text{test}} = 200$). Wilson 95% intervals are shown for binary metrics; cost is the observed implementation proxy, not total compute.

Metric	Observed value	95% CI	Unit
Test examples	200	–	count
Accuracy	**81.50**	[75.5, 86.3]	%
Coverage	**99.50**	[97.2, 99.9]	%
Mean relative cost	0.0794	–	proxy units
Total relative cost	15.89	–	proxy units

Model utilization. Table 2 shows that Qwen2.5-7B and Qwen3-32B account for ≈78% of the 206 attempts, while Llama-3.3-70B accounts for 16.5%. Dolphin-34B receives no traffic. These are realized policy traces on one split, not evidence that unused models are intrinsically ineffective; the larger split in Appendix B.6 shows similar concentration but does not establish stability.

Table 2: Per-model utilization on the primary split. Attempts can exceed 200 under escalation. Brackets are descriptive Wilson 95% intervals for attempt-level accuracy; green cells mark observed extrema, not significant differences.

Model	Attempts	Correct	Acc. [95% CI]	Output toks	Proxy
Qwen2.5-3B	5	1	20.0 [3.6, 62.4]	983	0.05
Phi-3.5-mini	5	3	60.0 [23.1, 88.2]	10	0.01
Llama-3.1-8B	1	0	0.0 [0.0, 79.3]	2	0.01
Qwen2.5-7B	17	16	**94.1** [73.0, 99.0]	2,167	2.10
Qwen3-32B	**144**	118	81.9 [74.9, 87.4]	288	0.55
Dolphin-34B	0	0	–	0	0.00
Llama-3.3-70B	34	25	73.5 [56.9, 85.4]	6,170	13.17
Total	**206**	**163**	–	**9,620**	**15.89**

4.3 Comparative Analysis

Figure 2 includes UCR (Dekoninck et al., 2025) and reference configurations associated with FrugalGPT (Chen et al., 2024) and RouterBench (Hu et al., 2024). At the plotted 81.5% operating points, CRISP records proxy cost 0.0794 versus 0.133 for UCR. The arithmetic difference is 40.3%, but it is *not* a reproducible practical-compute claim: the probe cost, input/output weights, latency, and complete baseline configurations are unavailable. CP-Router (Su et al., 2026), the closest conformal routing-only method, is discussed conceptually but was not rerun. Thus, this figure reports the available within-artifact comparison rather than establishing dominance over these systems.

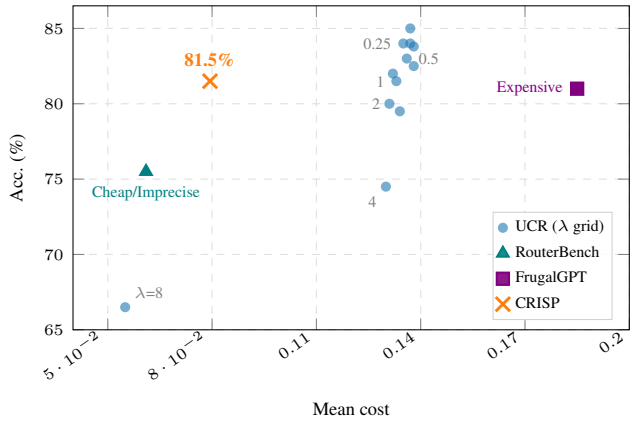


Figure 2: Descriptive proxy-cost comparison. UCR denotes the implemented unified routing-and-cascading grid; the points labeled RouterBench and FrugalGPT are reported reference configurations whose complete training and tuning specifications are unavailable in the artifact.

(a) Component ablation

Method	Acc. (%)	Mean cost
CRISP	**81.50**	0.0794
RoutingOnly	81.00	0.0758
CascadingOnly	54.00	0.0920
CheapestFirst	40.00	**0.0464**
RandomSingle	71.00	0.0968

(b) Gate-level sensitivity

$\alpha_r := \alpha_{\text{route}}, \quad \alpha_s := \alpha_{\text{stop}}$

α_r	α_s	Cost	Acc. (%)	Regime
0.05	0.10	0.0820	81.0	stricter stop
0.05	0.15	0.0794	**81.5**	main setting
0.05	0.20	0.0769	81.0	looser stop
0.10	0.15	0.0790	**81.5**	looser route
0.20	0.20	**0.0755**	80.5	loosest pair

Table 3: Primary-split diagnostics ($N_{\text{test}} = 200$). CascadingOnly changes multiple components and is not a clean stopping ablation. Green marks observed extrema and blue the prespecified setting; no paired or repeated-seed significance test is claimed.

Relative to RoutingOnly, CRISP obtains 163 rather than 162 correct answers and raises mean proxy cost from 0.0758 to 0.0794 (4.7%). The trace contains 206 attempts for 200 queries, hence only six additional model calls. Because paired correction/harm transitions were not retained, these totals do not show whether the stopping gate systematically rejects wrong answers without rejecting correct ones. Its demonstrated contribution here is therefore the fixed-model false-acceptance guarantee, not a statistically established accuracy gain.

4.4 Gate-Level Sensitivity and Coverage Semantics

Table 3(b) exposes the operating-point trade-off rather than presenting only the best row. Tightening the stopping level from 0.15 to 0.10 increases cost because fewer outputs pass and more queries escalate. Relaxing either gate reduces cost, but accuracy is not perfectly monotone because discrete ranks can alter the entire model sequence. The differences are within the sampling uncertainty of a 200-example test set, so the sweep is a sensitivity diagnostic rather than evidence that 0.15 is universally optimal.

The reported 99.5% coverage is protected-mode coverage: one query abstains because no eligible output passes, and that abstention counts as incorrect. Querying the largest model afterward is an explicitly unprotected fallback that must be reported separately; silently substituting it would void the gate interpretation.

The levels are not confidence probabilities for individual answers. Under exchangeability, they bound failure-pass rates for fixed model–task groups. Equation (23) gives the discrete resolution for small failure sets.

4.5 Finite-Sample Feasibility and Audit Scope

For lower-tail ranks, a gate can pass only if $m \geq \lceil 1/\alpha \rceil - 1$. With $N_{\text{cal}} = 40$, at most two task groups can reach 19 observations, so protected coverage on a balanced four-task test set is at most 50%; balanced calibration gives 10 examples per task and zero coverage. We therefore report no protected accuracy curve for this size. Table 4 also shows that uniform familywise routing at global level 0.05 is infeasible for both evaluated splits.

Table 4: Necessary failure-reference sizes for a nonempty protected gate. "Impossible" follows because $m_{k,t} \leq n_t$, independent of model accuracy.

Setting	n_t	$\alpha_{r,k}$	$m_{\min}$	Protected routing
Balanced small calibration	10	0.05	19	impossible
Primary, fixed-model level	50	0.05	19	possible if $m_{k,t} \geq 19$
Larger, fixed-model level	100	0.05	19	possible if $m_{k,t} \geq 19$
Primary, uniform familywise	50	$0.05/7$	139	impossible
Larger, uniform familywise	100	$0.05/7$	139	impossible

The paper consequently claims only fixed-model validity; it does not demonstrate a usable portfolio-wide protected operating point. A complete empirical audit must evaluate *every* candidate model on every held-out example and report, for each model–task pair, calibration failures, test failures, routing passes among test failures, stopping passes among test failures, and binomial intervals. Logging only cascade-reached candidates would condition on selection and would not test the theorem. Those sufficient statistics are absent from the reported logs, so the nominal in-distribution levels are not empirically audited here.

4.6 Larger-Split Robustness

On a disjoint $N_{\text{cal}} = N_{\text{test}} = 400$ split, CRISP records 86.0% accuracy $[82.3, 89.1]$, 99.5% coverage $[98.2, 99.9]$, and 0.1115 proxy cost (Appendix B.6). Because both partitions change, this is a second fixed-split result rather than evidence that more calibration data causes the improvement.

4.7 Task Performance and Distribution Shift

Task accuracy ranges from 78.0% on GSM8K to 86.0% on HellaSwag, with wide and overlapping intervals (Appendix B.5). Task-conditional normalization keeps ranks within task groups; it does not remove the need for a separate failure-reference set for each model–task pair.

Stress test under distribution shift. In GSM8K→MATH500 transfer (Lightman et al., 2023), the fraction of incorrect outputs passing a stopping gate targeted at 0.10 is 42.1% for an uncalibrated threshold and 14.3% for CRISP at $1.45\times$ proxy cost. CRISP still exceeds the target because exchangeability is broken; this illustrates Theorem A.4's degradation rather than providing OOD control or an in-distribution audit.

Mechanism-level evidence. Appendices B.7 and B.8 report that $\rho = 0.20$ ties $\rho = 0.05$ at 105/120 correct with lower proxy cost, while prompt edits change 5.5–13.0% of paths. Ordinary NLL, entropy alone, failure AUROC, and repeated trials remain untested; multiple-choice tasks use an option-margin score.

5 Conclusion

We introduced **CRISP**, a failure-conditional controller with finite-sample fixed-model routing and stopping guarantees under exchangeability. Rank resolution makes protected routing impossible under balanced $N_{\text{cal}} = 40$ and prevents uniform seven-model familywise routing at the evaluated sizes. The 200-query study is descriptive: CRISP returns 163 correct answers at 99.5% coverage, but gains only one answer over RoutingOnly at 4.7% higher proxy cost. Establishing a system-level advantage requires all-model failure-pass audits, paired stopping transitions, complete probe/token/latency accounting, ordinary-NLL controls, and matched repeated baselines.

Reproducibility statement

Sections 3–4 specify the scores, calibration rules, model portfolio, data splits, baselines, and metrics. Appendix A gives assumptions and proofs; Appendix C gives pseudocode; and Appendix D records hyperparameters, compute, and reporting details. Appendix E enumerates the artifacts intended for an anonymized review-time supplement and public release.

Ethics statement

The experiments use public benchmarks and existing open-weight models and do not involve human subjects or private data. Lowering inference cost may broaden access to useful LLM systems, but it may also reduce the cost of deploying unreliable or harmful applications. Appendix E discusses these risks and limits the claims to the evaluated control events and settings.

AI use statement

Generative AI tools were used to assist with manuscript editing, LaTeX formatting, and critical review of mathematical claims, proofs, and methodological presentation. They were not used to generate the reported experimental measurements. The authors are responsible for independently verifying all AI-assisted text, mathematics, citations, and artifacts and take responsibility for the final submission.

References

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

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. URL https://openreview.net/forum?id=cSimKw5p6R.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

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

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

Jasper Dekoninck, Maximilian Baader, and Martin T. Vechev. A unified approach to routing and cascading for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 12987–13010. PMLR, 2025.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid LLM: Cost-efficient and quality-aware query routing. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=02f3mUtqnM.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Eric Hartford, Lucas Atkins, Fernando Fernandes, and Cognitive Computations. Dolphin-2.9.1-yi-1.5-34b. Hugging Face model card, 2024. URL https://huggingface.co/dphn/dolphin-2.9.1-yi-1.5-34b.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. Routerbench: A benchmark for multi-LLM routing system. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2403.12031.

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

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. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 37733–37754. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1881. URL https://aclanthology.org/2026.findings-acl.1881/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Meta. Llama-3.3-70b-instruct model card. Hugging Face model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

Microsoft. Phi-3.5-mini-instruct model card. Hugging Face model card, 2024. URL https://huggingface.co/microsoft/Phi-3.5-mini-instruct.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36029–36047. PMLR, 2024. URL https://proceedings.mlr.press/v235/mohri24a.html.

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 from preference data. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html.

OpenAI. Api pricing. https://openai.com/api/pricing/, 2026. Accessed: 2026-09-23.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=pzUhfQ74c5. Poster.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. API is enough: Conformal prediction for large language models without logit-access. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 979–995. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.54. URL https://aclanthology.org/2024.findings-emnlp.54/.

Jiayuan Su, Fulin Lin, Zhaopeng Feng, Han Zheng, Teng Wang, Zhenyu Xiao, Xinlong Zhao, Zuozhu Liu, Lu Cheng, and Hongwei Wang. CP-Router: An uncertainty-aware router between LLM and LRM. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33065–33073, 2026. doi: 10.1609/aaai.v40i39.40589. URL https://ojs.aaai.org/index.php/AAAI/article/view/40589. Issue 39.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems*, 2025a. URL https://arxiv.org/abs/2506.01939.

Xinyuan Wang, Yanchi Liu, Wei Cheng, Xujiang Zhao, Zhengzhang Chen, Wenchao Yu, Yanjie Fu, and Haifeng Chen. MixLLM: Dynamic routing in mixed large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 10912–10922. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.naacl-long.545. URL https://aclanthology.org/2025.naacl-long.545/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01.ai. *arXiv preprint arXiv:2403.04652*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, 2019. doi: 10.18653/v1/P19-1472.

A Theoretical Guarantees and Proofs

This appendix states exactly what CRISP controls. The guarantees concern two failure events: selecting a model that is wrong on the current prompt and accepting an incorrect generated answer. They do not imply that the error rate conditional on acceptance is at most α; that quantity also depends on the base failure rate and the power of the gate on correct outputs.

A.1 Preliminaries

Let $Z_i = (X_i, Y_i, T_i)$ denote prompt, label, and task-type triples. For model M_k, let $\hat{y}_k(X_i)$ be its output under the same fixed decoding procedure used during calibration and deployment, define $L_{k,i} = \ell_k(X_i, Y_i) \in \{0, 1\}$, and write $L_k = \ell_k(X, Y)$ for a fresh test failure indicator. For a fixed model–task pair (k, t), the calibration failures are

$$\mathcal{F}_{k,t} = \{i \le n : T_i = t,\ L_{k,i} = 1\}, \qquad m_{k,t} = |\mathcal{F}_{k,t}|.$$

For scalar reference values $v_1, \dots, v_m$ and a test value v, recall the lower-tail rank

$$\widehat{p}(v; v_{1:m}) = \frac{1 + \sum_{j=1}^{m} \mathbb{1}[v_j \le v]}{m + 1}. \tag{24}$$

Exchangeability. The exact results require calibration and test examples, including any decoding randomness, to be exchangeable within task groups. Equivalently for each theorem below, conditional on $T = t$ and $L_k = 1$, the calibration failure scores and a fresh failure score must be exchangeable. A change in prompts, models, decoding, evaluation, or score definition can violate this condition.

Lemma A.1 (Lower-tail conformal rank)**.** *If $V_1, \dots, V_m, V_{m+1}$ are exchangeable, then for every $\alpha \in (0, 1)$,*

$$\mathbb{P}[\widehat{p}(V_{m+1}; V_{1:m}) \le \alpha] \le \alpha. \tag{25}$$

Proof. Without ties, Eq. (24) is the normalized ascending rank of V_{m+1} among $m + 1$ exchangeable values. At most $\lfloor \alpha(m + 1) \rfloor$ ranks satisfy the event. With ties, using $\le$ assigns the upper rank and can only make the event less likely. □

A.2 Routing: false-eligibility control

For calibration failure $i \in \mathcal{F}_{k,t}$, let $R_{k,i} = \tilde{r}(X_i; t)$. The test routing rank is

$$p^{\text{route}}_{k,t}(x) = \widehat{p}\big(\tilde{r}(x; t); \{R_{k,i}\}_{i \in \mathcal{F}_{k,t}}\big)\,. \tag{26}$$

Theorem A.2 (Finite-sample false-eligibility control)**.** *Fix model k, task t, and $m_{k,t} = m$. Under the exchangeability condition above,*

$$\mathbb{P}[k \in \mathcal{E}(X, t) \mid \ell_k(X, Y) = 1,\ T = t,\ m_{k,t} = m] \le \alpha_{\text{route},k}. \tag{27}$$

The inequality also holds after averaging over $m_{k,t}$.

Proof. Conditional on a fresh model failure, the m calibration failure scores and the fresh score are exchangeable. Apply Lemma A.1 to Eq. (26). □

The theorem controls the probability that an actually failing model enters the eligible set. It does not guarantee that every correct model is eligible or directly bound the error of the adaptively selected model. Smaller $\alpha_{\text{route},k}$ is stricter and may increase cost or abstention.

A.3 Interruption: false-acceptance control

For calibration failure $i \in \mathcal{F}_{k,t}$, let $S_{k,i} = \tilde{s}_k(X_i, \hat{y}_k(X_i); t)$. Define

$$p^{\text{stop}}_{k,t}(x, \hat{y}) = \widehat{p}\big(\tilde{s}_k(x, \hat{y}; t); \{S_{k,i}\}_{i \in \mathcal{F}_{k,t}}\big)\,. \tag{28}$$

Theorem A.3 (Finite-sample false-acceptance control)**.** *Fix model k, task t, and $m_{k,t} = m$. Under the exchangeability condition above,*

$$\mathbb{P}[A_k(X;t) = 1 \mid \ell_k(X,Y) = 1,\ T = t,\ m_{k,t} = m] \le \alpha_{\text{stop},k}. \tag{29}$$

The inequality also holds after averaging over $m_{k,t}$.

Proof. Conditional on model failure, the calibration output scores and a fresh output score are exchangeable. Apply Lemma A.1 to Eq. (28). □

Equation (29) bounds $\mathbb{P}(A_k = 1 \mid L_k = 1, T = t)$, the fraction of failures passing the gate. It is not the selective error $\mathbb{P}(L_k = 1 \mid A_k = 1, T = t)$, which also depends on the model's base failure rate and the probability that correct outputs pass.

A.4 Distribution shift

Let $P_{k,t}^{F,\text{cal}}$ and $P_{k,t}^{F,\text{test}}$ be the calibration and deployment distributions of either gate score conditional on $T = t$ and $L_k = 1$. For probability measures P, Q, define

$$d_{\text{TV}}(P,Q) = \sup_B |P(B) - Q(B)|, \qquad \Delta_{k,t} = d_{\text{TV}}\left(P_{k,t}^{F,\text{cal}}, P_{k,t}^{F,\text{test}}\right).$$

Let α_k denote the routing or stopping level for the gate under consideration.

Theorem A.4 (Degradation under failure-score shift)**.** *Let $\mathcal{A}(\mathcal{D}_{\text{cal}})$ be the score region passing either calibrated gate. If calibration failures are sampled from $P_{k,t}^{F,\text{cal}}$ and deployment failures from $P_{k,t}^{F,\text{test}}$, then*

$$\mathbb{E}_{\mathcal{D}_{\text{cal}}} P_{k,t}^{F,\text{test}}(\mathcal{A}(\mathcal{D}_{\text{cal}})) \le \alpha_k + \Delta_{k,t}, \tag{30}$$

Proof. For every fixed calibration set and measurable region $\mathcal{A}$, $P_{k,t}^{F,\text{test}}(\mathcal{A}) \le P_{k,t}^{F,\text{cal}}(\mathcal{A}) + \Delta_{k,t}$ by the definition of total variation. Average over the calibration set. The expected first term is at most α_k by Theorem A.2 or A.3 applied to a fresh draw from the calibration failure distribution. □

Thus, exact validity is not claimed under arbitrary shift. If deployment failures move toward unusually favorable scores, the pass rate may exceed its target by the conditional failure-score discrepancy.

A.5 Adaptive end-to-end control

Per-model validity does not automatically survive adaptive model selection. Let $k^\star(X) = \perp$ when no model is eligible and otherwise denote the first selected model. Let $\hat{k}(X) = \perp$ when the cascade abstains and otherwise denote the model whose output is returned. Allocate budgets satisfying

$$\sum_{k=1}^{K} \alpha_{\text{route},k} \le \alpha_{\text{route}}, \qquad \sum_{k=1}^{K} \alpha_{\text{stop},k} \le \alpha_{\text{stop}}.$$

Theorem A.5 (Familywise adaptive control)**.** *Under the assumptions of Theorems A.2 and A.3 for every model,*

$$\mathbb{P}\Big[(k^\star = \perp \ \vee\ \ell_{k^\star}(X,Y) = 0) \ \wedge\ (\hat{k} = \perp \ \vee\ \ell_{\hat{k}}(X,Y) = 0)\Big] \ge 1 - \alpha_{\text{route}} - \alpha_{\text{stop}}. \tag{31}$$

Proof. Because $k^\star$ is selected from the eligible set,

$$\{k^\star \ne \perp, \ell_{k^\star}(X,Y) = 1\} \subseteq \bigcup_{k=1}^{K} \{L_k = 1, k \in \mathcal{E}(X,T)\}.$$

After multiplying the conditional bound in Theorem A.2 by the model's failure probability and averaging over tasks, each event on the right has probability at most $\alpha_{\text{route},k}$. Similarly,

$$\{\hat{k} \ne \perp, \ell_{\hat{k}}(X,Y) = 1\} \subseteq \bigcup_{k=1}^{K} \{L_k = 1, A_k(X;T) = 1\},$$

and each term is at most $\alpha_{\text{stop},k}$. A union bound and the budget inequalities prove the result. □

The experiments use common per-model levels and therefore instantiate only the fixed-model guarantees. A deployment requiring Eq. (31) must allocate global budgets across models, but feasibility must be checked before deployment. For $K = 7$, uniform allocation of global levels $(0.05, 0.15)$ requires at least 139 routing failures and 46 stopping failures per model–task pair. Neither evaluated split can meet the routing requirement, so no nontrivial familywise protected operating point is demonstrated here.

B Additional Experiments

B.1 Oracle Diagnostic

Oracle Upper Bound (Cheapest-Correct). To contextualize CRISP's operating point, we also report an oracle upper bound that selects the cheapest model that answers correctly for each query (computed offline). This oracle achieves substantially higher accuracy and lower cost than any deployable policy, serving as a diagnostic on portfolio complementarity.

Table 5: Oracle upper bound (cheapest-correct) on the same evaluation split. This is not deployable, but indicates the attainable cost–accuracy region if perfect selection were possible.

Policy	Accuracy	Mean cost	Notes
Oracle (cheapest-correct)	87.0%	0.0058	offline per-query selection

B.2 Deployment Audit and Threats to Validity

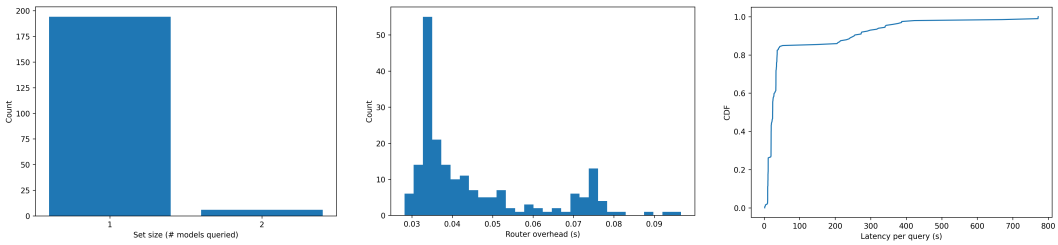
Table 6: Deployment audit separating validity-critical assumptions (orange) from efficiency-critical components (green).

Component	Failure mode	Consequence	Required action
Exchangeability	prompt, model, decoding, or evaluator drift	gate levels need not control deployment failures	version the full pipeline; recalibrate after a change
Failure references	too few model–task failures	coarse ranks, excess abstention, unstable traffic	enlarge calibration data or prespecify valid pooling
Adaptive selection	per-model levels reused as a global claim	multiplicity can exceed the stated portfolio budget	allocate model-wise levels that sum to a global budget
Score power	correct and incorrect outputs have similar scores	validity remains, but cost or abstention increases	inspect alignment and redesign the score before recalibration
Cost model	prices or output lengths change	routing may be inefficient; rank validity is unchanged	refresh cost and acceptance estimates

Exchangeability requires calibration and deployment failures to remain comparable within each model–task group; drift in prompts, models, decoding, parsers, or evaluators can invalidate the probability statement. Unlabeled rank, abstention, utilization, length, and score diagnostics can trigger an audit but cannot certify validity. Deployment should collect delayed labels on a prespecified sample, report failure-pass intervals, and recalibrate after pipeline changes. Theorem A.4 quantifies degradation by total variation when the conditional failure-score discrepancy can be estimated.

B.3 Retained Execution Traces

The aggregate stopping comparison is therefore narrow: CRISP gains one correct answer over RoutingOnly while making six extra model calls and increasing the token proxy by 4.7%. The reported logs do not contain the paired transition table needed to distinguish incorrect-to-correct recoveries, correct-answer rejections, and stopping-induced abstentions. Figure 3 supports the attempt counts but cannot establish a causal or practically significant latency benefit.



(a) Models queried per test case. (b) Recorded routing overhead. (c) End-to-end latency CDF.

Figure 3: Execution traces retained for the primary run. The first panel agrees with 206 attempts over 200 queries: 194 queries use one model and six use two. The timing panels are descriptive only because no matched baseline latency trace, probe identifier, or energy measurement is available.

B.4 Finite-Sample Feasibility and Required Failure Audit

For a failure-reference set of size m, the smallest attainable lower-tail rank is $1/(m+1)$. Hence a gate at level α can pass only if

$$m \geq m_{\min}(\alpha) := \left\lceil \frac{1}{\alpha} \right\rceil - 1.$$

This deterministic check precedes any accuracy analysis. At the reported fixed-model levels, routing requires $m \geq 19$ and stopping requires $m \geq 6$. Under a uniform seven-model allocation of global levels $(0.05, 0.15)$, the respective requirements become $m \geq 139$ and $m \geq 46$. The familywise routing allocation is therefore impossible with either 50 or 100 calibration examples per task.

At $N_{\text{cal}} = 40$, at most two task groups can contain 19 observations, so protected coverage on a balanced four-task test set cannot exceed 50%. With balanced calibration, every task has ten examples and coverage must be zero. Accordingly, no protected accuracy curve is reported at this size; any nonzero curve would imply pooled references, an alternative rank, or an unprotected fallback outside the stated procedure.

An empirical theorem audit requires one row per model–task pair containing (i) calibration failure count, (ii) held-out failure count, (iii) routing passes among those failures, (iv) stopping passes among those failures, and (v) uncertainty intervals for both pass rates. Every candidate model must be evaluated on the held-out examples; restricting the audit to candidates reached by the adaptive cascade conditions on selection. These sufficient statistics are not present in the reported logs, so the current experiments do not empirically verify the nominal in-distribution pass levels.

B.5 Per-Task Results

Table 7: Task accuracy and primary model usage on the stratified fixed split (50 test examples per task). Brackets are Wilson 95% intervals; no across-task significance comparison is implied.

Task	Domain	Correct	Acc. [95% CI]	Dominant model
HellaSwag	Commonsense	43/50	**86.0** [73.8, 93.0]	Qwen3-32B
WinoGrande	Reference	41/50	82.0 [69.2, 90.2]	Qwen3-32B
ARC-Chall.	Science	40/50	80.0 [67.0, 88.8]	*Mixed (Tier 2/3)*
GSM8K	Math	39/50	78.0 [64.8, 87.2]	Qwen2.5-7B

B.6 Robustness to Larger Calibration and Test Splits

To check whether the qualitative behavior persists with more data, we repeat the experiment using a larger disjoint split with $N_{\text{cal}} = 400$ and $N_{\text{test}} = 400$, corresponding to 200 examples per dataset across calibration and test. All hyperparameters, including $(\alpha_{\text{route}}, \alpha_{\text{stop}}) = (0.05, 0.15)$, are unchanged. Because both calibration and test samples differ from the primary split, this is a robustness check rather than a controlled estimate of the causal effect of calibration size.

Table 8 compares the two fixed splits. The higher accuracy point estimate may reflect both additional calibration data and different test examples, so we do not attribute it solely to calibration size.

Table 8: CRISP on two fixed disjoint splits. Brackets are Wilson 95% intervals; different examples preclude a paired significance test. Cost is the implementation proxy.

Split	N_{cal}	N_{test}	Acc. [95% CI]	Cov. [95% CI]	Cost
Primary	200	200	81.5 [75.5, 86.3]	99.5 [97.2, 99.9]	**0.0794**
Larger	400	400	**86.0** [82.3, 89.1]	99.5 [98.2, 99.9]	0.1115

Table 9 gives the utilization breakdown. Mid-sized models account for most attempts, while Llama-3.3-70B is invoked in 45 of 402 attempts. Several portfolio models are never selected; on this split, the eligibility and expected-cost policy therefore concentrates traffic rather than spreading it uniformly.

Table 9: Per-model utilization for the larger split. Totals exceed N_{test} because of escalation. Brackets give descriptive Wilson 95% intervals for attempt-level accuracy, not significance tests.

Model	Attempts	Correct	Acc. [95% CI]	S1	S2	Out toks	Cost
Qwen2.5-3B	1	1	100.0 [20.7, 100]	1	0	177	0.18
Phi-3.5-mini	0	0	–	0	0	0	0.00
Llama-3.1-8B	0	0	–	0	0	0	0.00
Qwen2.5-7B	57	55	96.5 [88.1, 99.0]	57	0	7,842	15.68
Qwen3-32B	**299**	247	82.6 [77.9, 86.5]	299	0	598	1.20
Dolphin-34B	0	0	–	0	0	0	0.00
Llama-3.3-70B	45	41	91.1 [79.3, 96.5]	43	2	6,881	27.52
Total	402	344	–	–	–	–	44.58

The larger split preserves the qualitative pattern of selective use of mid-sized models and occasional escalation. Multiple random splits are still needed to separate split variability from the effect of additional calibration data.

Table 10: **Ablation on the entropy-gating fraction ρ in EG-NLL.** We report overall evaluation accuracy and total cascade cost (token-based proxy) aggregated across four tasks. For each task, we use 60 validation examples, split into 30 calibration and 30 evaluation examples; totals are reported over $4 \times 30 = 120$ evaluation queries (ARC-Challenge, GSM8K, HellaSwag, WinoGrande XL). Green cells mark observed column extrema and blue marks the selected setting; this is a single validation split, not a repeated-seed significance analysis.

ρ	Accuracy (%)	Total cost
0.05	**87.50**	22.42
0.10	86.67	20.80
0.15	85.83	19.73
0.20	**87.50**	**15.43**
0.25	85.83	16.41
0.30	85.83	16.77

B.7 Effect of the entropy-gating fraction ρ.

We ablate the fraction ρ of highest-entropy forking tokens used by EG-NLL in a separate validation study, holding all other components fixed. Among the tested values, $\rho = 0.2$ ties the highest observed accuracy and records lower proxy cost than $\rho = 0.05$. This justifies fixing $\rho = 0.2$ for the reported runs, but it does not establish the proposed score's advantage: $\rho = 1$ (ordinary sequence-average NLL), entropy alone, and prediction of actual failures were not evaluated. Moreover, the three multiple-choice tasks use the option-margin output score, so this ablation is not evidence that EG-NLL drives their stopping decisions. The result is therefore a parameter-sensitivity check, not a comparative score study.

B.8 Prompt Sensitivity Ablation

To assess sensitivity to superficial prompt edits, we run an ablation on the *same* calibration and evaluation split used in the main experiments ($N_{\text{cal}} = 200$, $N_{\text{test}} = 200$). We apply semantics-preserving prefix, suffix, and whitespace perturbations, then execute the full pipeline with all hyperparameters, including failure-pass levels, fixed.

We report two complementary quantities. First, we report end-task **Accuracy**. Second, we report **Routing Stability**, defined as the fraction of evaluation examples for which the *cascade decision path* matches the baseline (i.e., the same initially selected model and the same stop-versus-escalate outcome).[1] Since this ablation uses the same split as the main table, the baseline accuracy is $\mathrm{Acc}_{\mathrm{base}} = 81.5\%$, and we additionally report $\Delta\mathrm{Acc}$ in percentage points relative to this baseline.

Table 11: Prompt sensitivity on the fixed primary split ($N_{\mathrm{cal}} = N_{\mathrm{test}} = 200$). $\Delta\mathrm{Acc}$ is relative to the unperturbed 81.5% baseline; stability is the fraction with the same routed model and stop/escalation decision. Green marks observed column extrema. Paired predictions are unavailable, so no significance claim is made.

Perturbation mode	Accuracy (%)	ΔAcc (pp)	Stability (%)
Prefix (neutral instruction)	**82.50**	**+1.00**	90.50
Suffix (neutral reminder)	82.00	+0.50	**94.50**
Whitespace (format-only edits)	**82.50**	**+1.00**	87.00

On this split, accuracy changes by $+0.5$ to $+1.0$ percentage points and routing stability ranges from 87.0% to 94.5%. Benign edits can therefore change the decision path even when aggregate accuracy is similar; broader perturbation sets and repeated splits are needed before claiming general prompt robustness.

C Algorithmic Details

We detail offline calibration and protected online inference.

Algorithm 1 CRISP calibration (failure-conditional split conformal)

Require: Calibration set $\mathcal{D}_{\mathrm{cal}}$, models $\{M_k\}_{k=1}^{K}$, probe P, levels $\{\alpha_{\mathrm{route},k}, \alpha_{\mathrm{stop},k}\}$.
1: Compute prompt scores $r(x_i)$ and task-conditional normalizers.
2: **for** $k = 1$ to K **do**
3: Generate $\hat{y}_k(x_i)$; compute losses $\ell_k(x_i, y_i)$ and output scores $s_k(x_i, \hat{y}_k(x_i))$.
4: **for** each task t **do**
5: Form $\mathcal{F}_{k,t}$ using Eq. (9).
6: Store normalized routing and stopping scores for $i \in \mathcal{F}_{k,t}$.
7: Estimate leave-one-out acceptance probabilities $\widehat{p}_{k,t}$, optionally by prompt-risk bin.
8: **end for**
9: **end for**
10: **return** normalizers, failure-reference scores, and acceptance estimates.

Labels are used offline to identify failures against which lower-tail ranks are computed. Different models generally fail on different examples, so their eligible regions are model-specific. If a model has no calibration failures for a task, protected mode assigns rank one and does not infer safety from the empty reference set.

Smaller α values are stricter: fewer failures can pass either gate, typically increasing escalation or abstention. Returning the largest model after all gates reject is possible as an explicitly unprotected fallback, but it is outside the protected coverage reported here.

D Reproducibility, Statistical Reporting, and Compute

Protocol. We evaluate seven open-weight LLMs on HellaSwag, WinoGrande XL, ARC-Challenge, and GSM8K. The primary evaluation uses 100 examples per dataset, split into $N_{\mathrm{cal}} = 200$ calibration and $N_{\mathrm{test}} = 200$ test examples. The larger experiment uses disjoint sets with $N_{\mathrm{cal}} = 400$ and $N_{\mathrm{test}} = 400$. Calibration labels identify model failures, construct task-conditional failure-reference sets, normalize scores, and estimate leave-one-out acceptance probabilities; final accuracy, coverage, and cost use held-out test examples only.

[1] We evaluate stability at the policy-decision level rather than token-level generation equality, since decoding randomness and minor formatting differences can change the generated text even when the routing and interruption decisions are identical.

Algorithm 2 CRISP protected inference

Require: Prompt x, task t, models $\{M_k\}$, failure-reference scores, acceptance estimates.
1: Compute $\tilde{r}(x;t)$ and $p^{\mathrm{route}}_{k,t}(x)$ for every model using Eq. (11).
2: Form $\mathcal{E}(x,t)$ using Eq. (16); **abstain** if $\mathcal{E}(x,t)=\emptyset$.
3: Order eligible models by cost and choose the suffix beginning at $k^\star$ via Eq. (20).
4: **for** each model k in the selected suffix **do**
5: Query M_k and compute $p^{\mathrm{stop}}_{k,t}(x,\hat{y}_k(x))$.
6: **if** $p^{\mathrm{stop}}_{k,t}(x,\hat{y}_k(x)) \le \alpha_{\mathrm{stop},k}$ **then**
7: **return** $\hat{y}_k(x)$
8: **end if**
9: **end for**
10: **abstain**

Portfolio and hyperparameters. The portfolio is Qwen2.5-3B, Phi-3.5-mini-instruct, Llama-3.1-8B, Qwen2.5-7B, Qwen3-32B, Dolphin-2.9.1-Yi-1.5-34B, and Llama-3.3-70B-Instruct. Unless stated otherwise, common per-model failure-pass levels are $(\alpha_{\mathrm{route}}, \alpha_{\mathrm{stop}}) = (0.05, 0.15)$, $\rho = 0.2$, and task-conditional normalization is used. The prompt score uses $c_{\mathrm{in}} = 0.0015$, $c_{\mathrm{dig}} = 0.04$, and $\delta_{\mathrm{code}} = 0.2$. These values are shared across datasets and are not tuned per task; ρ is selected in the separate validation study in Appendix B.7.

Metrics and uncertainty. Accuracy is the fraction of test examples for which the final returned answer matches ground truth, counting abstentions as incorrect; coverage is the non-abstention fraction. Relative cost is the implementation-specific token proxy in Eq. (21). The reported record does not include its numerical weights, probe identifier, or per-query ledger, so the values are not independently reproducible as total compute. Wilson 95% intervals accompany binary metrics. The fixed-model theorems are not empirically audited because model–task failure counts and all-model held-out pass events were not retained; the uniform familywise routing allocation is infeasible at the evaluated calibration sizes. We report a larger split, prompt perturbations, and a GSM8K→MATH500 stress test, but not repeated random splits.

Compute. Experiments ran on three NVIDIA RTX A6000 GPUs with 48 GB VRAM each. The 70B model used a multi-GPU configuration; smaller models used one GPU when memory permitted. Figure 3 preserves the available latency and routing-overhead traces, but no matched baseline timing or energy run is available. A reproducible release must add the exact probe, token weights, prefill/scoring treatment, baseline configurations, per-query ledger, software versions, and paired predictions.

E Limitations, Asset Use, and Responsible Release

Limitations. CRISP controls the probability that an actual failure passes a gate, conditional on task and model; it does not directly bound selective error. Exact validity requires exchangeable failure scores under an unchanged model, decoding rule, evaluator, and score. Small reference sets make ranks discrete, the familywise allocation is unusable at the present data sizes, and the in-distribution pass events are not logged. Empirically, the stopping ablation differs by one correct answer without paired transitions; ordinary NLL and CP-Router controls are absent; cost accounting is incomplete; and the study uses modest fixed splits, four reasoning benchmarks, and one portfolio. These omissions preclude claims of statistical significance, portfolio-wide protection, or practical compute savings.

Assets and release. The work uses public benchmarks and existing open-weight models cited in the paper. New assets are the calibration and inference pipeline, EG-NLL implementation, expected-cost policy, prompt templates, evaluation scripts, and plotting utilities; no new foundation model or scraped dataset is introduced. Any review-time release must be anonymized. The planned package includes model and dataset access instructions, preprocessing, licenses or terms-of-use notes where available, and instructions for reproducing the tables and figures; the final version will link the public repository and license.

Responsible use. Potential benefits include lower inference cost and auditable gate parameters. Potential harms include reducing the cost of deploying LLM systems where errors are consequential or insufficiently supervised. The guarantees do not justify high-stakes use without task-specific validation, shift monitoring, and an appropriate abstention or human-review policy.