CONFORMAL RISK-CONTROLLED ROUTING FOR LARGE LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

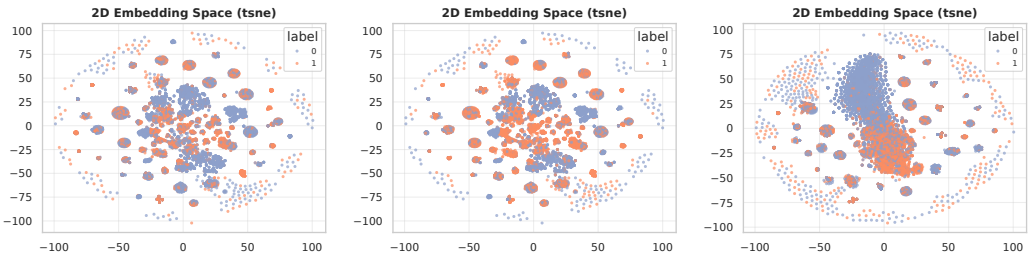
Recent advances in small-scale large language models have shown that compact models can successfully handle an expanding range of natural language and reasoning tasks. This progress opens the door to more affordable AI inference services by enabling broader use of cost-efficient models. However, existing approaches often fail to fully exploit small models due to fuzzy boundaries of their capabilities. In this paper, we propose a risk-controlled routing framework that dynamically selects among models of different scales, with a strong emphasis on maximizing the utility of smaller models. Our framework integrates supervised contrastive learning to enhance the separability of smaller-model capabilities and grounds its routing mechanism in conformal risk control, providing theoretical guarantees on system-level routing risk. Across benchmarks, our method delivers cost–accuracy performance that is comparable to or better than strong baselines, with an absolute accuracy improvement of $\sim 3.49\%$ at equal cost and up to $\sim 36\%$ cost reduction at comparable accuracy.

1 INTRODUCTION

Large language models (LLMs)(OpenAI, 2025a; DeepSeek-AI et al., 2025; Grattafiori et al., 2024) have progressed rapidly, demonstrating strong performance across a wide range of natural language and reasoning tasks. To increase accessibility, model families such as GPT(OpenAI, 2025b), Gemma(Team et al., 2025), and Qwen(Yang et al., 2025) are released in multiple scales, each with distinct accuracy-efficiency trade-offs. This diversity creates an opportunity to improve system-level efficiency: rather than relying exclusively on a single large model, queries can be adaptively routed to models of different scales. Realizing this potential requires solving a central systems problem: *LLM routing*(Ding et al., 2022; 2024; Hu et al., 2024). The goal is to design a mechanism that dynamically selects the most suitable model for each query, where suitability entails two criteria: achieving sufficient accuracy to solve the task and maintaining an inference cost affordable to most users.

Recent research on LLM routing mainly falls into two categories. The first is learning-based approaches, such as RouteLLM (Ong et al., 2025), HybridLLM (Ding et al., 2024), TO-router (Stripelis et al., 2024), BEST-route (Ding et al., 2025), and RouterDC (Chen et al., 2024b). The second is similarity-based approaches, where queries are embedded and routed based on their proximity or consistency in representation space, including clustering or nearest-neighbor retrieval (e.g., k-means–based partitioning in RouterBench (Hu et al., 2024)) and output-consistency methods such as Smoothie (Guha et al., 2024). These approaches do not require supervised training of a router but instead leverage the structural similarity among queries or the agreement among model outputs.

The primary limitation of current routing paradigms is their failure to fully exploit small, cost-efficient models, which are frequently bypassed even when capable. This under-utilization arises not from explicit design choices but from the inherent difficulty of predicting their performance. At its core lies a representation challenge: a small model's ability to correctly answer a query does not consistently align with its semantic representation. For example, two semantically similar queries may be mapped to nearby points in a standard embedding space, yet a small model may succeed on one and fail on the other (see Figure 1). The generic embeddings of prior work are insensitive to fine-grained differences in model capabilities, causing routers to be overly cautious and default to larger, more expensive models. Complementing this representational flaw is the risk-aware decision-



(a) (b) (c)

Figure 1: Embedding space separability (t-SNE). (a) Off-the-shelf embeddings: small-model correctness labels are heavily mixed. (b) Larger model: better but imperfect separation. (c) After SCL on the small model: clear delineation of answerable vs. unanswerable queries. Colors denote correctness (1/0).

making challenge. Assigning a complex query to an underpowered model wastes computation and may yield unexpected or incorrect responses, whereas sending a simple query to an expensive model incurs unnecessary cost. Existing methods typically rely on heuristic thresholds or fixed rules, but they lack a principled framework to formally quantify and control routing risk, leaving system-level behavior fragile and hard to guarantee.

To address these challenges, we propose *Conformal Risk-Controlled Routing* (CR^2), a framework that integrates capability-aware representation learning with principled risk control and cost-aware selection. Inspired by greedy algorithms, the first stage focuses on exploiting the utility of the smallest, most economical model. To tackle the core representation challenge, we employ supervised contrastive learning (SCL)(Khosla et al., 2020) to construct embeddings augmented with model-specific answerability, enabling the router to separate queries that are semantically similar but have different outcomes on the small model. Queries that cannot be confidently assigned to the small model are escalated to a second-stage router, which selects a candidate set of stronger models. To address the risk-aware decision-making challenge, we ground the routing process in the *Conformal Risk Control* (CRC). Specifically, we define a system-level risk function using candidate-set model-level false-positive rate (FPR) and calibrate a global candidate threshold under a held-out calibration set, providing formal guarantees for routing risk. Within the resulting candidate sets, a simple cost-aware rule selects the lowest-cost model, yielding a predictable and tunable accuracy–cost trade-off.

The main contributions of this work are summarized as follows:

- We propose CR^2, a two-stage routing framework that prioritizes the cost-efficient models. By leveraging supervised contrastive learning to refine embeddings, the router distinguishes semantically similar queries with divergent answerability, overcoming a key limitation of prior embedding-based methods.
- To the best of our knowledge, this is the first work to introduce CRC into LLM routing. By defining a bounded, composite routing loss and calibrating a global candidate threshold, CR^2 provides formal guarantees that the expected risk is provably bounded below specified level α, while also yielding improved performance.
- Extensive experiments demonstrate that CR^2 establishes a new state of the art in LLM routing. It achieves an absolute accuracy improvement of approximately 3.49% (6% relative) over strong baselines such as EmbedLLM and single largest model, while simultaneously reducing overall operational cost.

2 Related Work

2.1 Model Routing in LLMs

Dynamic routing for efficiency spans multiple granularities: token-level mixtures-of-experts within a single forward pass (Fedus et al., 2022; Zhou et al., 2022; Li et al., 2025) and window-level

schemes such as speculative decoding (Leviathan et al., 2023; Lu et al., 2023; Chen et al., 2024c; Li et al., 2024). This work focuses on query-level routing, where an entire request is dispatched to one model from a pool. Existing methods include pre-generation routers that train lightweight selectors to pick a single model before inference (Ong et al., 2025; Ding et al., 2024; Feng et al., 2025; Stripelis et al., 2024; Ding et al., 2025) and post-generation cascades that escalate from cheaper to more expensive models until a quality criterion is met (Chen et al., 2024a). While effective, these approaches typically rely on fixed thresholds or heuristics and provide no distribution-free guarantees. Our framework complements this line by combining hierarchical routing with conformal risk control.

2.2 Capability-Aware Representations

Routing often hinges on representations that anticipate which model can solve a query. Early approaches embed models via accuracy profiles or simple classifiers to separate "easy" from "hard" queries (Zhuang et al., 2025; Ding et al., 2024). More recent work leverages contrastive objectives, either by jointly embedding queries and models (Chen et al., 2024b) or by modeling query–LLM relationships through transformer-based backbones (Jin et al., 2025). Other methods enrich embeddings with auxiliary signals, such as capability instructions that combine past performance and user prompts (Zhang et al., 2025b), or document-level context to capture knowledge shifts (Zhang et al., 2025a). However, these embeddings can conflate semantic similarity with *answerability*, leading to under-utilization of smaller, cost-efficient models. Our approach instead applies supervised contrastive learning to shape embeddings so that proximity reflects model-specific answerability, improving small-model utilization without sacrificing accuracy.

2.3 Risk-Aware Decision Making and Conformal Prediction

Beyond representation, routing is also a risk management problem. Conformal prediction provides distribution-free reliability guarantees, but classical coverage does not directly address cost–accuracy trade-offs. Conformal Risk Control (CRC) extends these tools to general bounded risks with finite-sample guarantees (Angelopoulos et al., 2024), and has been applied to mitigate hallucination in single-LLM settings (Overman et al., 2024; Chen et al., 2025). Other recent conformal methods include CP-Router, which uses uncertainty estimates for routing (Su et al., 2025), and another that optimizes risk and prediction set size during training (Noorani et al., 2024). To our knowledge, we are the first to introduce CRC into the routing pipeline itself: we calibrate a global candidate threshold so that the expected system-level routing risk—whose Stage-2 component is the candidate-set model-level false-positive rate—remains within a user-specified tolerance, while a cost-aware selector realizes efficiency gains.

3 Preliminaries

3.1 Problem Formulation

We study routing over a pool of large language models (LLMs) with heterogeneous sizes and inference costs. Let $\mathbb{Q}$ denote the space of queries and $\mathbb{M} = \{M_1, \ldots, M_K\}$ the available models. Each model M_i is associated with an inference cost $c_i > 0$ and a correctness indicator$A_i(q) = \mathbf{1}[\, M_i(q) = y \,]$, where $M_i(q)$ is the output of M_i on query q and y is the ground-truth answer; hence $A_i(q) \in \{0, 1\}$ indicates whether M_i answers q correctly. A routing strategy is a mapping $R : \mathbb{Q} \to \mathbb{M}$ that assigns a model $M_{R(q)}$ to each query q. The system-level correctness on q is $A_{R(q)}(q)$. We evaluate a routing strategy R by its expected accuracy

$$\mathrm{Acc}(R) = \mathbb{E}_{q\sim\mathbb{Q}}\big[A_{R(q)}(q)\big]\,, \tag{1}$$

and its expected cost

$$\mathrm{Cost}(R) = \mathbb{E}_{q\sim\mathbb{Q}}\big[c_{R(q)}\big]\,. \tag{2}$$

The routing problem is thus a multi-objective optimization that maximizes accuracy while minimizing cost:

$$\max_{R}\ \big(\,\mathrm{Acc}(R),\ -\mathrm{Cost}(R)\,\big), \tag{3}$$

equivalently $\min_R \big(1 - \mathrm{Acc}(R), \mathrm{Cost}(R)\big)$, which induces a Pareto frontier.

Remark. In Section §4 we instantiate R via a two-stage architecture that prioritizes the smallest model when safe and escalates otherwise; here we only establish notation and objectives.

3.2 Conformal Risk Control

CRC (Angelopoulos et al., 2024) is a statistical framework that generalizes classical conformal prediction from coverage guarantees to controlling the expectation of a general loss function. Given a base predictor, a calibration set $\{(X_i, Y_i)\}_{i=1}^n$, and a user-specified risk level $\alpha \in (0, 1)$, CRC provides a recipe for calibrating a parameter $\hat{\lambda}$ to ensure that the expected loss on a new test point does not exceed α.

The framework operates on a family of predictors $C_\lambda(x)$ indexed by a parameter $\lambda \in \Lambda$. This parameter λ controls the **conservativeness** of the predictor's output. We define a loss for each calibration example as

$$L_i(\lambda) = \ell(C_\lambda(X_i), Y_i)\,. \tag{4}$$

A critical requirement of the framework is that the loss function ℓ must be **monotone non-increasing** with respect to λ. This ensures that a more conservative choice of λ will not result in a higher loss. This property holds for many useful applications, such as controlling the false negative rate in multilabel classification or token-level F1 loss in question answering (Angelopoulos et al., 2024).

The goal of CRC is to select a data-driven threshold $\hat{\lambda}$ such that the following expected risk guarantee holds for a new test point (X_{n+1}, Y_{n+1}):

$$\mathbb{E}\Big[L_{n+1}(\hat{\lambda})\Big] \le \alpha. \tag{5}$$

CRC achieves this by calculating the empirical risk $\widehat{\mathcal{R}}(\lambda) = \frac{1}{n}\sum_i L_i(\lambda)$ on the calibration set and finding the least conservative λ that satisfies a high-probability risk bound. For a loss bounded by B, this is typically:

$$\hat{\lambda} = \inf\left\{ \lambda \in \Lambda \;\middle|\; \frac{n}{n+1}\,\widehat{\mathcal{R}}(\lambda) \;+\; \frac{B}{n+1} \;\le\; \alpha \right\}. \tag{6}$$

This guarantee is distribution-free and holds for finite samples. When ℓ is chosen as the miscoverage loss, $\ell(C_\lambda(X), Y) = \mathbf{1}\{Y \notin C_\lambda(X)\}$, CRC reduces exactly to classical conformal prediction.

4 Methodology

In this section, we introduce Conformal Risk-Controlled Routing, a framework designed to address the dual challenges of representation and risk-aware decision-making in LLM routing. The core of our approach is to decompose the global routing task into two specialized sub-problems: (1) a high-precision binary prediction for the single, most cost-effective model, and (2) a multi-label prediction to identify capable models from the remaining expert pool. Crucially, instead of combining these stages with fragile heuristics, we unify them under the CRC framework. This is achieved by designing a global risk function and a corresponding decision algorithm, which together provide a provable guarantee that the system-level trade-off between cost and accuracy is explicitly controlled.

4.1 Problem Decomposition

We parameterize a hierarchical router by thresholds $\theta = (t_1, t_2)$:

$$R_\theta(q) = \begin{cases} M_1, & \text{if } s_1(q) \ge t_1, \\ \text{Select}\big(\mathcal{C}_{t_2}(q)\big), & \text{otherwise}, \end{cases} \tag{7}$$

where $s_1(q) \in [0, 1]$ estimates the success probability of M_1 on q, $\mathcal{C}_{t_2}(q) = \{\, i \ge 2 : \hat{p}_i(q) \ge t_2 \,\}$ is a candidate set among larger models, and $\text{Select}(\cdot)$ is a cost-aware rule. Section 4.2 describes how to obtain $s_1(q)$; Section 4.3 details $\{\hat{p}_i(q)\}_{i\ge 2}$ and the CRC calibration of θ.

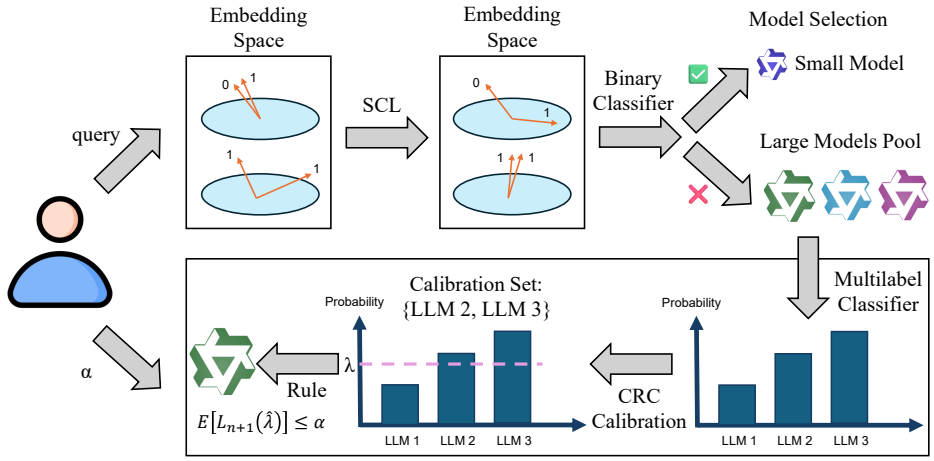


Figure 2: Overview of the CR^2 framework. A binary classifier, operating on a capability-aware embedding space shaped by SCL, first attempts to route a query to the small model. If the query is rejected, a multilabel classifier scores the expert model pool. Finally, CRC uses these scores to calibrate a decision threshold that guarantees the final cost-optimal selection adheres to a user-specified risk tolerance.

4.2 Capability-Aware Filtering

The first stage of CR^2 constructs a high-precision filter for the smallest model M_1, which predicts the binary correctness label $A_1(q) \in \{0, 1\}$ for a given query q. This is achieved through a two-phase procedure: (i) fine-tuning a text encoder to learn capability-aware representations; and (ii) training a classification head on these embeddings.

Architecture. The module consists of a pretrained encoder g_θ, a projection head u_φ, and a classification head h_ψ. To enrich the embedding for contrastive learning, u_φ is designed as an *attention pooling projector*, which uses learnable query vectors to attend to token-level encoder outputs and yield more informative representations than mean pooling. The classification head is a two-layer MLP.

Phase 1: Representation Learning via SCL. We use SCL (Khosla et al., 2020) to endow the Stage-1 filter with a representation whose geometry reflects the *answerability* of the smallest model M_1, rather than mere semantic similarity.

Setup. Let g_θ be a pretrained text encoder and u_φ a projection head (we use an attention-pooling projector). Given a query q, we form a normalized embedding

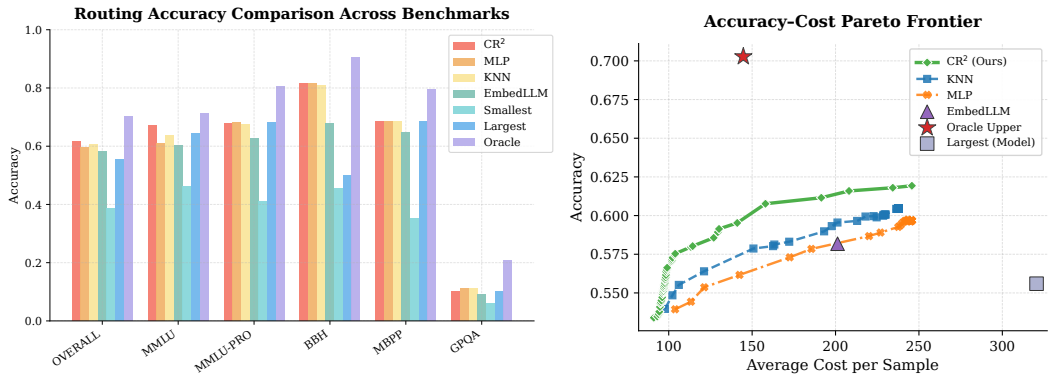
$$\boldsymbol{z}_q = \frac{u_\varphi\big(g_\theta(q)\big)}{\big\|u_\varphi\big(g_\theta(q)\big)\big\|_2}. \tag{8}$$

For a minibatch $\{(q_i, y_i)\}_{i=1}^B$, labels are $y_i = A_1(q_i) \in \{0, 1\}$, where $A_1(q)$ indicates whether M_1 answers q correctly (cf. Preliminaries).

Supervised contrastive loss. With temperature $\tau > 0$, the per-anchor SCL loss is

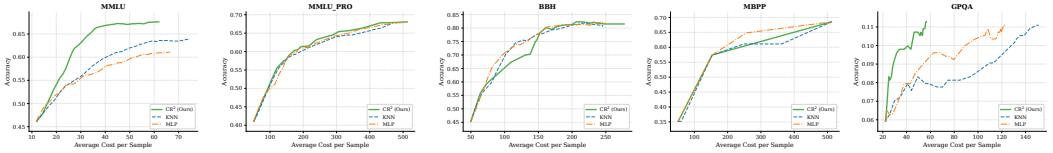
$$\mathcal{L}_i^{\mathrm{SCL}} = -\frac{1}{|\mathbb{P}(i)|} \sum_{p \in \mathbb{P}(i)} \log \frac{\exp\big(\boldsymbol{z}_i^\top \boldsymbol{z}_p/\tau\big)}{\sum_{a \in \mathbb{A}(i)} \exp\big(\boldsymbol{z}_i^\top \boldsymbol{z}_a/\tau\big)}, \tag{9}$$

where $\mathbb{P}(i) = \{\, p \neq i : \; y_p = y_i \,\}$ is the set of positives for anchor i and $\mathbb{A}(i) = \{\, a \neq i \,\}$ the set of all non-anchor samples in the batch (anchors with $|P(i)| = 0$ are skipped). Minimizing $\mathcal{L}^{\mathrm{SCL}} =$



(a) (b)

Figure 3: (a) Routing accuracy of CR^2 compared to baselines. CR^2 router performs better almost across the whole test set. (b) Accuracy–cost trade-off of different routing strategies, where CR^2 achieves superior Pareto efficiency compared to baselines.



(a) (b) (c) (d) (e)

Figure 4: Accuracy–cost trade-off of different routing strategies per benchmark.

$\sum_i \mathcal{L}_i^{\text{SCL}}$ pulls together queries that M_1 handles similarly (both solvable or both unsolvable) and pushes apart those with different outcomes, thereby reshaping the embedding space to be capability-aware with respect to M_1.

Classifier training on capability-aware embeddings. After SCL fine-tuning, we *freeze* the encoder g_θ and projector u_φ, and train a lightweight classification head h_ψ on the capability-aware representations. The head is trained to predict the success of model M_1, i.e., the binary label $y = A_1(q) \in \{0, 1\}$. It outputs two logits, and we convert their difference into a probability estimate via the sigmoid function:

$$s_1(q) \;=\; \sigma\big(\ell_1(q) - \ell_0(q)\big) \;\approx\; \Pr\big[A_1(q) = 1\big]. \tag{10}$$

Since the smallest model correctly handles only a limited proportion of queries, the training data for its success predictor suffers from a natural class imbalance. To mitigate this, the head is optimized by minimizing a class-weighted binary cross-entropy loss, where weights are determined by the inverse frequency of each class. At inference, the Stage-1 router routes to M_1 when $s_1(q) \geq t_1$ and otherwise escalates. We treat t_1 as a fixed gate (set on a held-out set) and use CRC (§4.3) to calibrate the Stage-2 candidate threshold so that the overall system-level routing risk satisfies the specified budget.

4.3 Multilabel Classification and CRC Calibration

For queries deferred by Stage 1 (i.e., $s_1(q) < t_1$), a multilabel head scores the remaining models

$$\hat{\mathbf{p}}(q) = \big(\hat{p}_2(q), \ldots, \hat{p}_K(q)\big) \in [0, 1]^{K-1}, \tag{11}$$

where $\hat{p}_i(q)$ estimates the probability that M_i answers q correctly. Let $y_{ij} = A_i(q_j) \in \{0, 1\}$ denote the ground-truth outcome for model M_i on query q_j. Given a global threshold $\lambda \in [0, 1]$, we define the candidate set

$$\mathcal{C}_\lambda(q) = \{\, i \in \{2, \ldots, K\} :\; \hat{p}_i(q) \geq \lambda \,\}. \tag{12}$$

Our goal is to select λ with a distribution-free, finite-sample risk guarantee *for the entire routed system* under a fixed gate t_1.

Per-query loss and monotonicity. On a held-out calibration set $\mathcal{D}_{\text{cal}} = \{(q_j, \mathbf{y}_j)\}_{j=1}^{n}$ (assumed exchangeable with test data), we define a bounded per-query loss

$$L_j(\lambda) = \begin{cases} 1 - y_{1j}, & \text{if } s_1(q_j) \geq t_1, \\ \dfrac{|\{\, i \in \mathcal{C}_\lambda(q_j) :\ y_{ij} = 0 \,\}|}{\max\big(1,\ |\{\, i \geq 2 :\ y_{ij} = 0 \,\}|\big)}, & \text{if } s_1(q_j) < t_1, \end{cases} \tag{13}$$

i.e., a misclassification indicator when routed to M_1, and the *model-level false-positive rate* within the candidate set otherwise. By construction $L_j(\lambda) \in [0, 1]$ and, holding t_1 fixed, $L_j(\lambda)$ is *non-increasing* in λ (larger λ shrinks $\mathcal{C}_\lambda$ and cannot add false positives). Hence the empirical risk

$$\widehat{\mathcal{R}}(\lambda) = \frac{1}{n} \sum_{j=1}^{n} L_j(\lambda) \tag{14}$$

is also non-increasing in λ. **We provide a formal proof that our composite loss function in Eq. equation 13 satisfies the crucial monotonicity property required by CRC in Appendix B.**

Calibrating λ via conformal risk control. We apply CRC for bounded losses (here $B = 1$). For a user-specified tolerance $\alpha \in [0, 1]$, CRC selects

$$\lambda^* = \inf \left\{ \lambda \in [0, 1] :\ \underbrace{\frac{n}{n+1}\, \widehat{\mathcal{R}}(\lambda) \ +\ \frac{1}{n+1}}_{\text{CRC upper bound on } \mathbb{E}[L(\lambda)]} \leq \alpha \right\}. \tag{15}$$

Choosing the *smallest* feasible λ^* maximizes candidate-set size under the same risk budget, preserving downstream cost opportunities while maintaining the distribution-free, finite-sample guarantee $\Pr\big(\mathbb{E}[L(\lambda^*)] \leq \alpha\big) \geq 1 - \delta$.

Final selection rule and fallback. At test time we use the fixed gate t_1 and set $t_2 = \lambda^*$. For $s_1(q) \geq t_1$, route to M_1; otherwise select

$$\text{Select}\big(\mathcal{C}_{t_2}(q)\big) = \arg\min_{i \in \mathcal{C}_{t_2}(q)} c_i \quad \text{(ties broken by larger } \hat{p}_i(q)\text{)}. \tag{16}$$

If $\mathcal{C}_{t_2}(q) = \varnothing$, we fall back to $\arg\max_{i \geq 2} \hat{p}_i(q)$ or a pre-specified robust model (see ablations). This policy, together with equation 15, yields distribution-free, finite-sample control of the expected composite risk in equation 13.

5 Experiment Results

5.1 Experimental Setup

We conduct a comprehensive set of experiments to evaluate the performance of our proposed method.

Model Pool and Costs. Our experiments utilize a pool of widely-used, open-source LLMs from Qwen3 family (Yang et al., 2025), which provides a realistic spectrum of capabilities and inference costs. Our model pool $\mathbb{M} = \{\text{Qwen3-1.7B}, \text{Qwen3-4B}, \text{Qwen3-8B}, \text{Qwen3-14B}\}$. We define the inference cost for each model based on the total number of input tokens according to official API price, normalizing them relative to the largest model Qwen3-14B. The relative costs are 0.15, 0.3, 0.5, 1.0 for Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Qwen3-14B, respectively.

Datasets. Following the reproducible protocol of EmbedLLM (Zhuang et al., 2025), we evaluate on a diverse query corpus spanning six challenging benchmarks covering expert knowledge, multi-step reasoning, and coding: MMLU (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024), GSM8K (Cobbe et al., 2021), Big-Bench Hard (BBH) (Suzgun et al., 2022), GPQA (Rein et al., 2023), and MBPP (Austin et al., 2021). To generate the ground-truth data, for each query q from these benchmarks and each model $M_i \in \mathbb{M}$, we run inference using the lm-evaluation-harness (Gao et al., 2024).

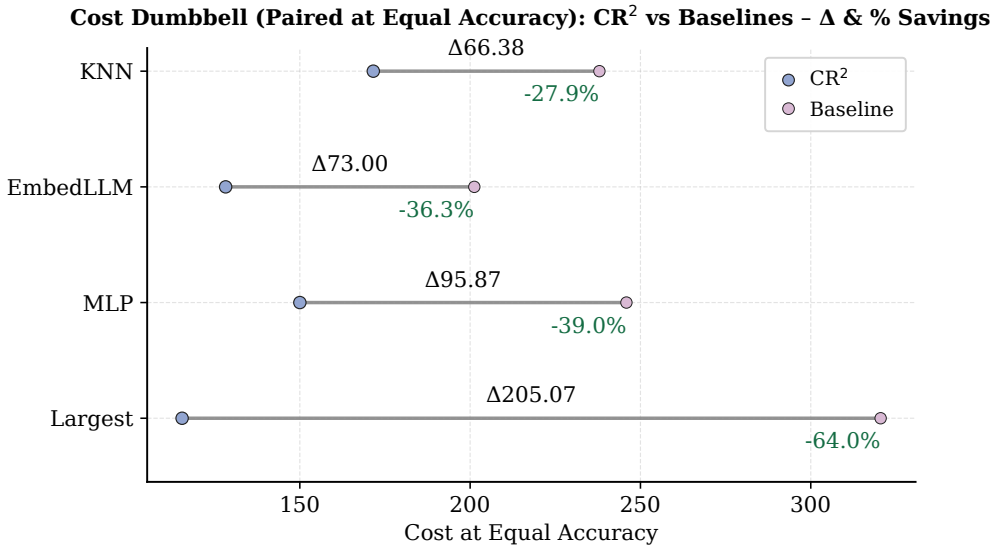


Figure 5: Cost dumbbell comparison at equal accuracy. CR^2 consistently achieves lower inference cost than baselines, with relative savings exceeding 60%.

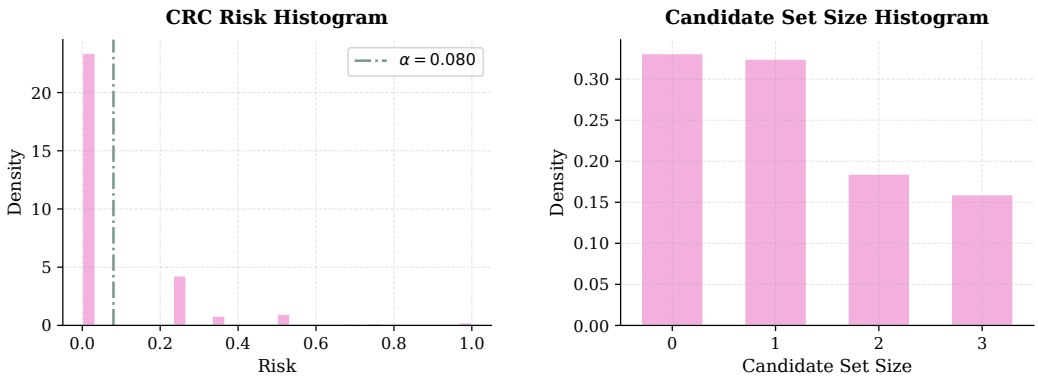
Baselines. We compare our method against a comprehensive set of baselines to rigorously evaluate its performance:

- **EmbedLLM** (Zhuang et al., 2025): The current state-of-the-art learning-based router, which uses general-purpose embeddings to predict model performance.
- **Always-Smallest**: A simple heuristic that always routes to the cheapest capable model.
- **Always-Largest**: A heuristic that always routes to the most largest model.
- **Oracle**: A theoretical upper bound that assumes perfect knowledge of each model's answerability for every query. It always selects the cheapest model that is known to answer the query correctly, defining the Pareto frontier.
- **MLP**: A non-hierarchical baseline where a MLP is trained on top of general-purpose sentence embeddings. It acts as a multi-class classifier to select a single model from the pool based on the highest output score.
- **KNN** (Zhuang et al., 2025): A non-parametric baseline that performs nearest-neighbor voting over query–model correctness outcomes. Each model is implicitly represented by its historical correctness tuples, and for a new query, the classifier predicts performance based on the majority vote of its nearest neighbors. We refer to this approach as KNN throughout the text.

Evaluation Metrics and Implementation Details. We evaluate all methods on two primary metrics: Routing Accuracy (%) and Average Per-Sample Token Cost (Avg. Cost). For cost, we normalize API prices to that of the most expensive model, compute each sample's token cost as the normalized price-per-token times its total input tokens, and then average over all samples. The ideal method should achieve high accuracy at a low cost. For our method, we use a pretrained `all-MiniLM-L6-v2`(Wang et al., 2020) as the base sentence encoder, which is then fine-tuned using the supervised contrastive loss.

CRC is configured to control the expected system-level risk, ensuring it remains below the user-specified tolerance α. For the main SOTA comparison, we set this risk level to $\alpha = 0.08$, though a broader analysis with varying α is also presented.

For each benchmark, we generate labels on its official training set, and then partition this labeled data into 80%/10%/10% splits for training, validating, and testing our router, respectively. Further implementation details, including all hyperparameters, are provided in Appendix A.



(a) (b)

Figure 6: Analysis of CR^2 routing. (a) Histogram of per-query risks under the calibrated CRC threshold, showing that the majority of samples lie well below the specified tolerance α. (b) Histogram of candidate set sizes produced by the second-stage router, illustrating the distribution of model subsets considered at inference.

5.2 Main Results

We now present the main experimental results, which show that CR^2 consistently outperforms strong baselines by achieving higher accuracy at equal cost and significantly reducing cost at equal accuracy.

First, in terms of routing accuracy, CR^2 achieves the best aggregate performance across five benchmarks (61.7%), with top or tied results on most individual tasks. As shown in Figure 3a, on BBH it surpasses the EmbedLLM baseline by 13.7 points, underscoring its strength on complex reasoning, while on MMLU it improves over the KNN baseline by 3.1 points, demonstrating robustness on knowledge-intensive evaluations. These results indicate that CR^2 narrows the gap to the oracle while preserving efficiency advantages over single-model deployments.

Second, when examining the accuracy–cost trade-off, CR^2 consistently defines the Pareto frontier. As shown in Figure 3b, its curve lies above all baselines and single-model settings, achieving higher accuracy at any given cost budget and substantially lower cost at a fixed accuracy. This highlights its ability to leverage smaller models effectively without sacrificing end-task performance.

Finally, we analyze the behavior of the risk calibration mechanism at inference time. Figure 6a shows that the calibrated router tightly controls the system's risk: per-query values are concentrated near zero, well below the user-specified tolerance of $\alpha = 0.08$. Figure 6b further illustrates how efficiency arises: for more than 65% of inputs, the router confidently selects a single candidate model (or none at all), while adaptively expanding to 2–3 candidates only on harder queries. This adaptivity explains how CR^2 remains both efficient and reliable in practice.

5.3 Ablation Studies

We toggle each component while holding others fixed and report routing accuracy and average per-sample token cost in Table 1. Enabling the two-stage design improves accuracy from **58.49%** to **60.74%** (+2.25 pts) and reduces cost from 223.10 to 202.18 (∼9.4%). Adding SCL lifts accuracy from **56.11%** to **58.34%** (+2.23 pts) and lowers cost from 201.56 to 196.04 (∼2.7%), indicating clearer separability of answerable vs. unanswerable queries for the small model. CRC calibration trims cost from 239.71 to 220.47 (∼8.0%) with essentially unchanged accuracy (61.05% vs. 60.97%).

Overall, the components are complementary: two-stage routing yields the largest gains, SCL sharpens the Stage-1 decision boundary, and CRC delivers reliable cost reductions under a risk budget.

6 CONCLUSION

In this work, we introduce CR^2, a novel hierarchical routing framework that learns capability-aware representations via supervised contrastive learning and, in a first for this domain, utilizes CRC to provide provable guarantees on the cost-accuracy trade-off. Experiments demonstrate that CR^2 establishes a new state-of-the-art, significantly improving both accuracy and cost-efficiency over strong baselines. By making the deployment of diverse LLMs more reliable and economically viable, our work represents a concrete step toward the affordable AI.

7 ETHIC STATEMENT

Our routing system could exacerbate fairness and bias issues if queries about sensitive topics are sent to smaller models that have not received the same safety alignment as larger models. While our framework does not inherently introduce bias, fairness depends on the quality and tuning of the model pool. Future work should explore routing criteria that explicitly account for fairness and safety.

8 REPRODUCIBLITY STATEMENT

We provide sufficient information to facilitate the reproduction of our results. The core code will be included in the supplementary material. Detailed implementation specifics, including model architectures, training procedures and hyperparameters, are described in the Appendix A. The full code will be publicly released on GitHub upon acceptance of the paper.

REFERENCES

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

Catherine Chen, Jingyan Shen, Zhun Deng, and Lihua Lei. Conformal tail risk control for large language model alignment. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=H8DkMvWnSQ.

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*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=cSimKw5p6R.

Shuhao Chen, Weisen Jiang, Baijiong Lin, James Kwok, and Yu Zhang. Routerdc: Query-based router by dual contrastive learning for assembling large language models. *Advances in Neural Information Processing Systems*, 37:66305–66328, 2024b.

Zhuoming Chen, Avner May, Ruslan Svirschevski, Yuhsun Huang, Max Ryabinin, Zhihao Jia, and Beidi Chen. Sequoia: Scalable, robust, and hardware-aware speculative decoding. *arXiv preprint arXiv:2402.12374*, 2024c.

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, 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, Chenyu Zhang, Chong Ruan,

Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, 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, 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, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, 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, Yanhong Xu, 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: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Dujian Ding, Sihem Amer-Yahia, and Laks V. S. Lakshmanan. On efficient approximate queries over machine learning models. *CoRR*, abs/2206.02845, 2022. URL https://doi.org/10.48550/arXiv.2206.02845.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=02f3mUtqnM.

Dujian Ding, Ankur Mallick, Shaokun Zhang, Chi Wang, Daniel Madrigal, Mirian Del Carmen Hipolito Garcia, Menglin Xia, Laks VS Lakshmanan, Qingyun Wu, and Victor Rühle. Bestroute: Adaptive llm routing with test-time optimal compute. *arXiv preprint arXiv:2506.22716*, 2025.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Tao Feng, Yanzhen Shen, and Jiaxuan You. Graphrouter: A graph-based router for LLM selections. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eU39PDsZtT.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius,

Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell,

Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Neel Guha, Mayee Chen, Trevor Chow, Ishan Khare, and Christopher Re. Smoothie: Label free language model routing. *Advances in Neural Information Processing Systems*, 37:127645–127672, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

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. In *Agentic Markets Workshop at ICML 2024*, 2024. URL https://openreview.net/forum?id=IVXmV8Uxwh.

Ruihan Jin, Pengpeng Shao, Zhengqi Wen, Jinyang Wu, Mingkuan Feng, Shuai Zhang, and Jianhua Tao. Radialrouter: Structured representation for efficient and robust large language models routing. *arXiv preprint arXiv:2506.03880*, 2025.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. EAGLE: Speculative sampling requires rethinking feature uncertainty. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=1NdN7eXyb4.

Zhongyang Li, Ziyue Li, and Tianyi Zhou. R2-t2: Re-routing in test-time for multimodal mixture-of-experts. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=oqPcOMafOF.

Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. Routing to the expert: Efficient reward-guided ensemble of large language models, 2023. *URL https://arxiv. org/abs/2311.08692*, 2023.

Sima Noorani, Orlando Romero, Nicolo Dal Fabbro, Hamed Hassani, and George J Pappas. Conformal risk minimization with variance reduction. *arXiv preprint arXiv:2411.01696*, 2024.

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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

OpenAI. Gpt-5 system card. https://cdn.openai.com/pdf/8124a3ce-ab78-4f06-96eb-49ea29ffb52f/gpt5-system-card-aug7.pdf, 2025a. Accessed: 2025-09-24.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025b. URL https://arxiv.org/abs/2508.10925.

William Overman, Jacqueline Jil Vallon, and Mohsen Bayati. Aligning model properties via conformal risk control. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=9OHXQybMZB.

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, 2023. URL https://arxiv.org/abs/2311.12022.

Dimitris Stripelis, Zijian Hu, Jipeng Zhang, Zhaozhuo Xu, Alay Dilipbhai Shah, Han Jin, Yuhang Yao, Salman Avestimehr, and Chaoyang He. Tensoropera router: A multi-model router for efficient llm inference. *arXiv preprint arXiv:2408.12320*, 2024.

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. *arXiv preprint arXiv:2505.19970*, 2025.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael

Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=y10DM6R2r3.

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.

Jiarui Zhang, Xiangyu Liu, Yong Hu, Chaoyue Niu, Fan Wu, and Guihai Chen. Query routing for retrieval-augmented language models. *arXiv preprint arXiv:2505.23052*, 2025a.

Yi-Kai Zhang, De-Chuan Zhan, and Han-Jia Ye. Capability instruction tuning: A new paradigm for dynamic llm routing. *arXiv preprint arXiv:2502.17282*, 2025b.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

Richard Zhuang, Tianhao Wu, Zhaojin Wen, Andrew Li, Jiantao Jiao, and Kannan Ramchandran. EmbedLLM: Learning compact representations of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Fs9EabmQrJ.

A Appendix 1: Training Implementation Details

A.1 Stage-1: Capability-Aware Filtering

SCL We initialize the text encoder with `sentence-transformers/all-MiniLM-L6-v2` and its tokenizer (max sequence length 512; EOS used as `pad_token` when absent). On top of the encoder we add an attention-pooling projector with two learned queries followed by a linear layer to a 384-d embedding; outputs are L2-normalized. Mini-batches contain 256 examples and are composed with a class-balanced sampler to stabilize SCL. We fine-tune the encoder and projector for 15 epochs using AdamW (LR 1×10^{-4} on projector/encoder, weight decay 0.01), cosine annealing with warm restarts ($\eta_{\min} = 1 \times 10^{-6}$), mixed precision on GPU (bfloat16), and gradient-norm clipping on the projector (max-norm 5.0). All runs use seed 42 and Weights&Biases for logging.

Binary classifier training on frozen embeddings. After SCL, both encoder and projector are frozen. We train a lightweight MLP head (LayerNorm $\rightarrow$ Linear($2d$) $\rightarrow$ GELU $\rightarrow$ Dropout(0.1) $\rightarrow$ Linear(2)) for 10 epochs with AdamW (LR 1×10^{-3}), the same cosine scheduler, and gradient clipping (max-norm 1.0). Class imbalance is handled via inverse-frequency class weights. This gate is fixed at deployment.

A.2 Stage-2: Multilabel Classification and CRC Calibration

Multilabel Classifier Training. We train a multilabel classifier to score the remaining models $\{M_i\}_{i=2}^{K}$ using a frozen MiniLM encoder and a lightweight head. Concretely, we instantiate `sentence-transformers/all-MiniLM-L6-v2` and freeze all backbone parameters. Inputs are tokenized with the corresponding tokenizer (padding enabled; `max_len`=512; EOS used as `pad_token` if absent). The head is an MLP (LayerNorm $\rightarrow$ Dropout(0.1) $\rightarrow$ Linear(d, $4d$) $\rightarrow$ GELU $\rightarrow$ Dropout(0.1) $\rightarrow$ Linear($4d$, $K-1$)), trained with `BCEWithLogitsLoss`. Batches contain 64 examples per GPU; we run distributed data-parallel training on 4 NVIDIA 4090 GPUs via `torchrun`, yielding an effective global batch size of 256. Optimization uses AdamW (LR $= 1\times10^{-3}$, weight decay $= 0.01$) for 20 epochs with a cosine schedule and 3% warmup. Gradients are clipped at 1.0.

To mitigate label imbalance across the $(K-1)$ binary targets, we use inverse-frequency weighted sampling, where per-example weights are the clipped ($[0.2, 5.0]$) sum of inverse per-class positive rates. Validation runs every epoch with distributed aggregation. Unless otherwise specified, we set the random seed to 42 and use 4 dataloader workers per process. The resulting probabilities serve as inputs to the CRC calibration step that determines the stage-2 candidate threshold used at deployment.

B Appendix 2: Proof of Monotonicity for the Composite Loss Function in CRC

Here, we formally prove that the composite loss function $L_j(\lambda)$ defined in Equation equation 13 of the main text is monotone non-increasing with respect to the threshold $\lambda \in [0, 1]$. This property is a prerequisite for the application of the Conformal Risk Control framework.

Proposition 1. *The composite loss function $L_j(\lambda)$ is monotone non-increasing with respect to λ.*

Proof. To prove that $L_j(\lambda)$ is monotone non-increasing, we must show that for any pair of thresholds $0 \leq \lambda_1 < \lambda_2 \leq 1$, it holds that $L_j(\lambda_2) \leq L_j(\lambda_1)$. The loss function is defined piece-wise based on the routing decision for a given query q_j, so we analyze each case.

Case 1: The query is handled by Stage 1 ($s_1(q_j) \geq t_1$). In this case, the loss is defined as $L_j(\lambda) = 1 - y_{1j}$. This value is a constant with respect to λ, as it does not depend on the threshold. A constant function is, by definition, monotone non-increasing. Thus, $L_j(\lambda_2) = L_j(\lambda_1)$, and the condition is satisfied.

Case 2: The query is handled by Stage 2 ($s_1(q_j) < t_1$). In this case, the loss is the model-level FPR:

$$L_j(\lambda) = \frac{|\{\, i \in \mathcal{C}_\lambda(q_j) :\ y_{ij} = 0 \,\}|}{\max\big(1,\ |\{\, i \geq 2 :\ y_{ij} = 0 \,\}|\big)}.$$

Let us analyze the components of this fraction. The denominator, $D = \max\big(1,\ |\{\, i \geq 2 :\ y_{ij} = 0 \,\}|\big)$, is a positive constant for a given query q_j, as it depends only on the ground-truth outcomes, not on λ.

The numerator, $N(\lambda) = |\{\, i \in \mathcal{C}_\lambda(q_j) :\ y_{ij} = 0 \,\}|$, is the number of incorrect models included in the candidate set. To prove that $L_j(\lambda)$ is non-increasing, it is sufficient to prove that the numerator $N(\lambda)$ is non-increasing.

The candidate set is defined as $\mathcal{C}_\lambda(q_j) = \{\, i \in \{2, \ldots, K\} :\ \hat{p}_i(q_j) \geq \lambda \,\}$. Consider our two thresholds such that $0 \leq \lambda_1 < \lambda_2 \leq 1$. For any model i to be in the set $\mathcal{C}_{\lambda_2}(q_j)$, its score must satisfy $\hat{p}_i(q_j) \geq \lambda_2$. Because $\lambda_2 > \lambda_1$, this condition implies that $\hat{p}_i(q_j) > \lambda_1$, which in turn means that model i must also be a member of the set $\mathcal{C}_{\lambda_1}(q_j)$.

Therefore, the candidate set at the higher threshold is a subset of the candidate set at the lower threshold:

$$\mathcal{C}_{\lambda_2}(q_j) \subseteq \mathcal{C}_{\lambda_1}(q_j).$$

The numerator $N(\lambda)$ counts the number of incorrect models within the candidate set. Let $I_{\text{incorrect}} = \{i \geq 2 : y_{ij} = 0\}$ be the set of all incorrect models for query q_j. The numerator can be written as $N(\lambda) = |\mathcal{C}_\lambda(q_j) \cap I_{\text{incorrect}}|$.

Since $\mathcal{C}_{\lambda_2}(q_j)$ is a subset of $\mathcal{C}_{\lambda_1}(q_j)$, the intersection of this smaller set with $I_{\text{incorrect}}$ must also be a subset of the intersection of the larger set with $I_{\text{incorrect}}$:

$$\mathcal{C}_{\lambda_2}(q_j) \cap I_{\text{incorrect}} \subseteq \mathcal{C}_{\lambda_1}(q_j) \cap I_{\text{incorrect}}.$$

The cardinality of a subset cannot be greater than the cardinality of the set that contains it. Thus, it follows that $N(\lambda_2) \leq N(\lambda_1)$. As the denominator is a positive constant, we have shown that the loss is non-increasing for this case as well.

Conclusion. Since the loss is monotone non-increasing in both cases, the composite loss function $L_j(\lambda)$ is proven to be monotone non-increasing with respect to λ over its entire domain. □

C Appendix 3: Ablation Studies

Table 1: Ablation study of Two Stage Routing, SCL and CRC.

	Accuracy (%)	**Avg. Cost**
w/o Two Stage Routing	58.49	223.10
w/ Two Stage Routing	60.74	202.18
w/o SCL	56.11	201.56
w/ SCL	58.34	196.04
w/o CRC	61.05	239.71
w/ CRC	60.97	220.47

D Appendix 4: The Use of Large Language Models

We used OpenAI ChatGPT and Google Gemini (Deep Research) strictly as productivity aids. Concretely, they were used to (i) polish wording and improve stylistic clarity of drafts, and (ii) help scope the literature at the project outset by suggesting search terms and candidate papers. No parts of the methods or results were generated by LLMs. Every citation surfaced during scoping was manually verified against primary sources, and no model-generated references were accepted. No confidential data were shared with the tools. The authors take full responsibility for the content of this paper.

Table 2: Sensitivity of CRC to the calibration sample size ($\alpha = 0.1$). The risk remains close to α while accuracy shows only small variations.

Calibration size N_c	λ	Acc	Avg token cost	Mean risk	Mean set size
100	0.9053	0.6035	**191.651**	0.1090	**1.4267**
250	0.9286	**0.6082**	206.417	**0.0907**	1.2233
500	0.9199	0.6069	201.863	0.0979	1.3107
1000	0.9266	0.6074	202.217	0.0931	1.2523

Table 3: CRC performance under different α values (refit). The observed risk closely matches the target α.

Variant	α	λ	Acc	Avg token cost	Mean risk	Mean set size
CRC@α	0.05	0.9827	**0.6095**	233.213	0.0520	0.5640
CRC@α	0.10	0.9253	0.6074	202.176	**0.0949**	1.2781
CRC@α	0.15	0.8712	0.5952	**150.947**	0.1497	**1.7532**

E Appendix 5: Out-of-Distribution Evaluation

Fig. 7 shows the accuracy–cost Pareto frontier evaluated on a combined out-of-distribution test set constructed from PIQA and ARC-Easy, which the router was not exposed to during training. Each point represents the average inference cost per sample under a specific routing configuration. The results reflect the model pool's performance and routing behavior in a distribution-shift setting rather than in-distribution generalization.

Our method CR^2 (green diamonds) forms a competitive Pareto frontier relative to KNN and MLP routers across the evaluated cost range. The Largest Model baseline (grey square) provides an upper-cost reference, while Oracle Upper Bound (red star) denotes the idealized maximum achievable accuracy if the best model were chosen per sample. The frontier of EmbedLLM is also plotted for comparison.

Overall, this OOD evaluation illustrates that routing continues to extract favorable accuracy–cost trade-offs even when the input distribution differs from training, and that the achievable frontier may exceed the standalone performance of the largest model.

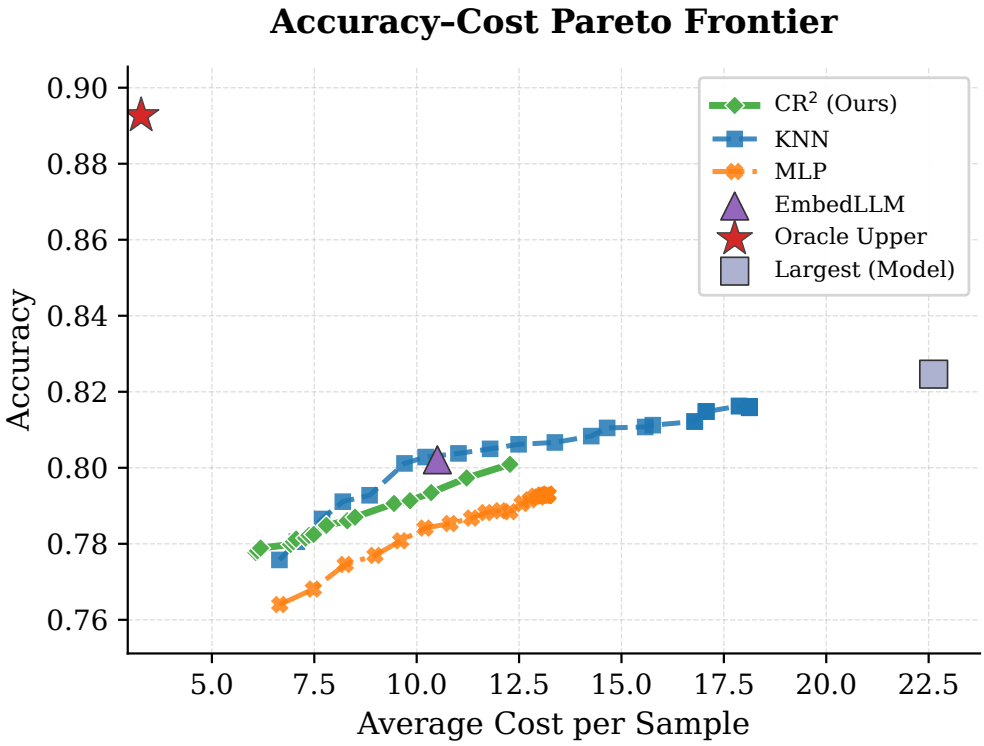


Figure 7: Accuracy–cost trade-off of different routing strategie under OOD settings.