Risk-Aware Decision-Theoretic Routing for Reliable Use of Large Language Models

Lujin Zhao*†, Sujuan Qin*†, Yijie Shi*†
*State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing, China
†School of Cyberspace Security,
Beijing University of Posts and Telecommunications, Beijing, China

***Abstract*—Routing queries to large language models (LLMs) is a knowledge-driven decision-making problem under uncertainty. Most existing routing approaches optimize expected performance and implicitly adopt a uniform failure cost assumption, which does not hold in safety-critical settings. We formulate instance-level LLM routing from a decision-theoretic perspective and cast it as a cost-aware optimization problem targeting tail-risk control. We introduce lightweight and interpretable risk abstractions that integrate performance prediction, predictive uncertainty, and extrapolation uncertainty, enabling principled risk estimation prior to model invocation. Based on these abstractions, we develop a one-shot decision-theoretic routing framework that balances reliability and inference cost without static thresholds or multi-model cascading. Experiments across reasoning-intensive, professional, and safety-critical benchmarks demonstrate that the proposed approach substantially improves reliability on high-risk and out-of-distribution queries while maintaining competitive inference efficiency. By providing interpretable decision-relevant signals, our framework supports robust and explainable model selection for mission-critical LLM deployment.**

***Index Terms*—large language model routing, risk-aware decision making, conditional value at risk, one-shot model selection, uncertainty quantification**

I. Introduction

Large language models (LLMs) are increasingly deployed in real-world applications ranging from routine question answering to mission-critical decision support [1]. In such deployments, system reliability is determined not only by average performance, but also by the consequences of rare yet high-impact failures. This naturally frames LLM deployment as a decision-making problem under uncertainty, where model selection constitutes an action and failures incur heterogeneous, context-dependent costs. In this setting, the router acts as a lightweight knowledge layer that mediates between user queries and a heterogeneous model pool, reasoning about latent risk before committing to inference, e.g., selecting between lightweight assistants for casual queries and high-reliability models for medical or legal decision support.

Although high-capacity models generally provide stronger reliability, their substantial inference cost makes uniform deployment impractical. To balance quality and efficiency, model routing has emerged as a key paradigm for dynamically assigning queries to appropriate models [2]. Most existing approaches optimize expected performance using preference predictors or classifier-based gating [3], [4]. Such methods implicitly assume uniform failure consequences, overlooking the fact that real-world queries exhibit highly heterogeneous risk profiles: errors in casual interactions are often tolerable, whereas failures in legal, financial, or safety-critical scenarios may incur severe consequences [5].

Corresponding author: Sujuan Qin (qsujuan@bupt.edu.cn)

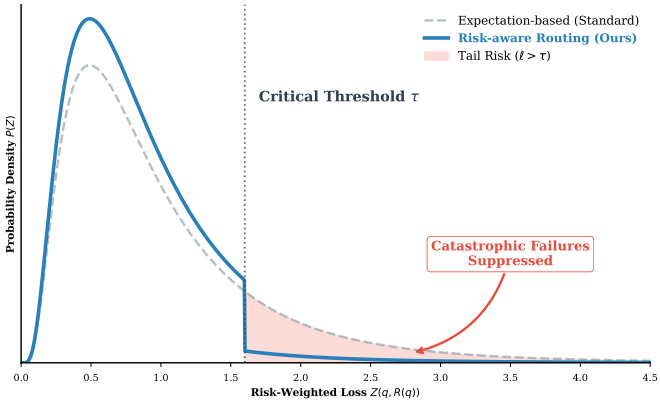

Fig. 1. Conceptual comparison of expectation-based and risk-aware routing. Risk-aware routing explicitly controls tail risk, suppressing rare but high-impact failures.

As illustrated in Fig. 1, expectation-based routing tends to produce a heavy tail of high-impact failures. This motivates a shift toward risk-aware, decision-theoretic routing that explicitly prioritizes reliability on high-stakes queries [6], [7]. A central challenge, however, is that failure costs and tail-risk statistics are unobservable at inference time. Routing decisions must be made prior to model invocation and under potential distributional shifts [8], rendering conventional performance predictors insufficient for identifying queries prone to rare but catastrophic failures.

To address this challenge, we introduce a set of lightweight, instance-level risk abstractions that serve as a symbolic–numeric knowledge representation of latent risk exposure. These abstractions integrate performance prediction with predictive and extrapolation uncertainty, providing decision-relevant signals for an adaptive one-shot routing policy. Unlike static thresholding or multi-model cascading [9], our approach

operates under a strict single-call constraint with negligible overhead.

Our main contributions are summarized as follows:

- We formulate LLM routing as an instance-level, risk-aware decision-theoretic problem that targets tail-risk reliability under inference budget constraints.
- We propose a unified set of lightweight risk abstractions as a knowledge representation of latent failure risk, combining performance prediction with predictive and extrapolation uncertainty.
- We design an adaptive one-shot routing policy that enables transparent and explainable decision making without relying on static thresholds or multi-model cascading.
- We demonstrate that the proposed framework substantially suppresses catastrophic failures while maintaining competitive efficiency across diverse benchmarks, providing a robust foundation for reliable LLM deployment. To support reproducibility, the source code and datasets used in this study will be made publicly available upon acceptance.

II. Related Work

Model routing for LLMs can be viewed as a knowledge-driven decision-making problem over heterogeneous model capabilities. Existing approaches primarily focus on expectation-based optimization, heuristic cascading, or uncertainty-aware routing under implicitly risk-neutral objectives, with limited treatment of heterogeneous failure consequences.

Preference-Driven and Utility-Based Routing. A large body of work formulates routing as a one-shot selection guided by predicted utility or cost–performance trade-offs, including RouteLLM [2], HybridLLM [3], MixLLM [10], and Causal LLM Routing [11]. Recent extensions incorporate safety-related signals, such as task-level safety classification in SafeRoute [12] or domain-specific risk awareness in RsynLLM [13], but these methods largely optimize expected outcomes and do not explicitly model instance-level risk heterogeneity or tail failure severity within a general decision-theoretic framework.

Cascading and Deferral-Based Routing. Sequential and deferral-based strategies, including FrugalGPT [14], AutoMix [9], and Semantic Cascades [15], improve average performance through adaptive multi-stage invocation. However, such approaches incur additional latency and inference cost, and rely on heuristic escalation rules, limiting their applicability in latency-sensitive or strict one-shot deployment settings.

Uncertainty-Aware Routing and Confidence Estimation. Another line of work leverages uncertainty signals derived from predictive entropy, calibration, or self-reported confidence [16], motivating routers such as RouterDC [17], CP-Router [18], and CARGO [19]. While effective for reducing average error, these methods primarily treat uncertainty as a proxy for correctness and do not distinguish routine low-confidence cases from rare but high-impact failures under distributional shift.

Risk-Aware Decision Making and Tail-Risk Control. Risk-sensitive reasoning has been extensively studied in decision theory [20], with coherent risk measures such as conditional value at risk (CVaR) enabling formal control of tail losses [6]. Recent work applies these principles to LLM systems, including Conformal Arbitrage [21] and risk-aware trustworthy generation or alignment settings [7], [22]. In contrast to prior approaches, our work brings risk-aware decision theory to the *meta-level* of model routing, enabling principled tail-risk-aware routing under a strict one-shot constraint without cascading or static thresholds.

III. Methods

As illustrated in Fig. 2, we propose a risk-aware one-shot routing framework grounded in decision theory. For each incoming query, the router infers lightweight epistemic signals capturing performance expectations, predictive uncertainty, and distributional shift, and evaluates them over a hierarchical model space to select the model tier in a single step, prioritizing tail-risk control under inference budget constraints.

A. Problem Formulation

a) Hierarchical Model Space and One-Shot Constraint.: Let Q denote the query space. The system maintains a pool of models organized into an ordered hierarchy of L tiers, denoted by $\mathcal{H} = \{1, \dots, L\}$, where higher tiers correspond to greater capability and reliability. We assume a monotonic cost–performance trade-off such that for any $l_j > l_i$, the inference costs satisfy $c(l_j) > c(l_i)$. Under a strict one-shot constraint, the routing policy $R : Q \to \{1, \dots, L\}$ is a deterministic mapping. Once a tier $l = R(q)$ is selected, the corresponding model is invoked to generate a response, incurring an inference cost $c(l)$ and a latent response loss $\ell(q, l)$. No intermediate feedback or multi-stage escalation is permitted.

b) Risk-Weighted Loss and Heterogeneous Stakes.: To capture heterogeneous failure consequences, we introduce a failure severity weight $w(q) \geq 0$, which serves as a prior abstraction of the query's stakes. High-stakes queries (e.g., legal, medical, or safety-critical) are assigned larger weights than casual or low-impact queries. Importantly, $w(q)$ is independent of the observed outcome at inference time and can be derived from domain metadata, query classification, or policy rules. We define the risk-weighted loss as

$$Z(q, l) = w(q) \cdot \ell(q, l), \tag{1}$$

which jointly reflects failure likelihood and severity. In practice, $w(q)$ can be assigned using lightweight query classifiers (e.g., keyword-based rules or compact text encoders) with negligible additional latency.

B. Risk-Aware Routing Objective

Most existing routing approaches optimize expected loss $\mathrm{E}[\ell(q, R(q))]$ under cost constraints. In decision-critical settings, however, system reliability is often dominated by rare but high-impact failures. To explicitly control this tail behavior, we adopt CVaR as the primary reliability objective.

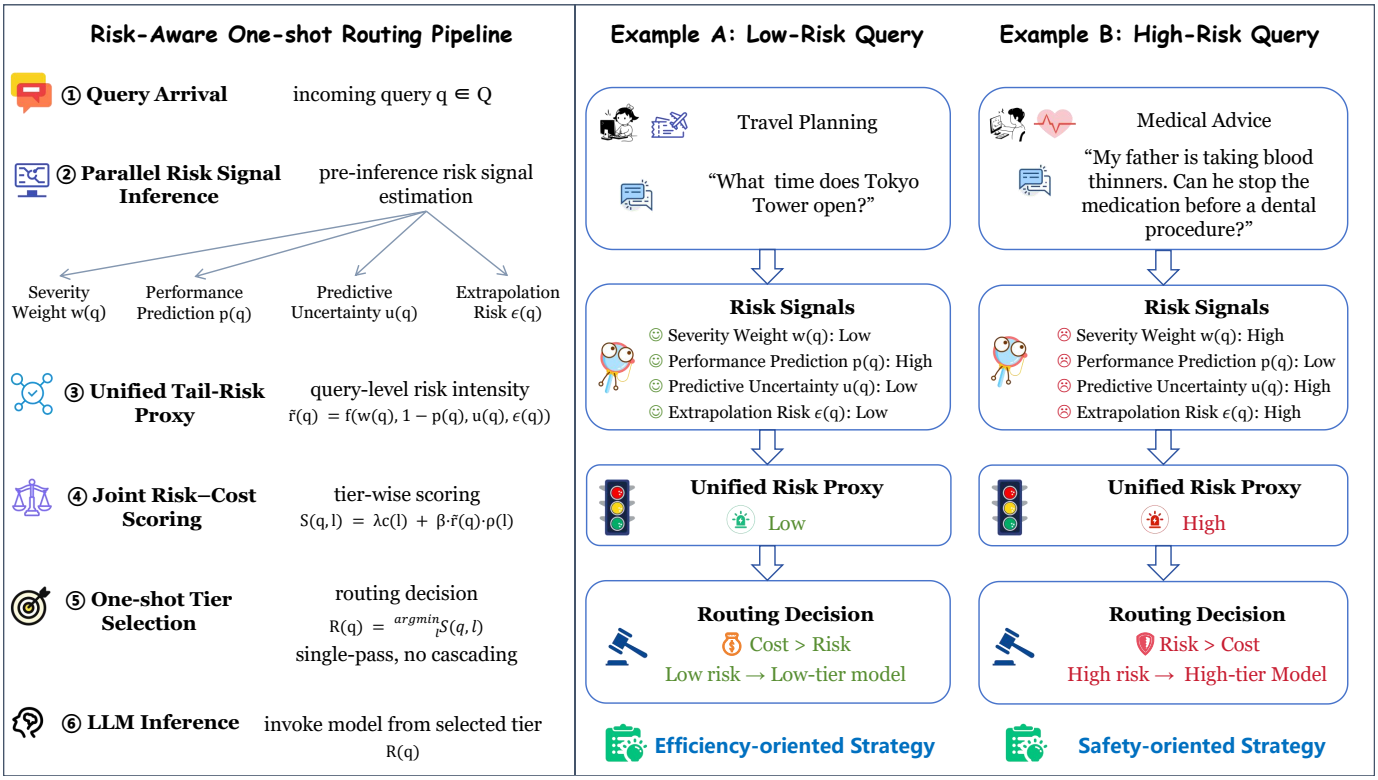

Fig. 2. Decision-Theoretic Architecture of the Risk-Aware One-Shot Routing Framework.

For a confidence level $\alpha \in (0,1)$, the CVaR of a routing policy R is defined as

$$\mathrm{CVaR}_\alpha(R) = \mathrm{E}[Z(q,R(q)) \,|\, Z(q,R(q)) \geq \mathrm{VaR}_\alpha(R)]\,, \quad (2)$$

where

$$\mathrm{VaR}_\alpha(R) = \inf\{\zeta : \Pr(Z(q,R(q)) \leq \zeta) \geq \alpha\}\,. \quad (3)$$

We impose an expected inference budget constraint $\mathrm{E}[c(R(q))] \leq B$. Direct optimization of CVaR under this constraint is intractable in the one-shot setting, as the loss distribution is unobservable at inference time. We therefore optimize a Lagrangian relaxation of the constrained problem using a cost-sensitive risk surrogate.

C. Epistemic Risk Abstractions

We interpret risk estimation as introspective epistemic reasoning, in which the router assesses the limits of its own knowledge before committing to a routing decision. To this end, we construct three complementary epistemic signals.

a) Performance Prediction.: Let $x = \phi(q)$ denote the embedding of query q. The router estimates the adequacy of the baseline tier ($l = 1$) via

$$p(q) = \sigma(g_\theta(x))\,, \quad (4)$$

where g_θ is a lightweight prediction head. Lower values of $p(q)$ indicate a greater need for higher-capability models.

b) Predictive Uncertainty.: Predictive uncertainty $u(q)$ is quantified using entropy or margin statistics derived from the router's internal classifier, reflecting instability in its belief.

c) Extrapolation Uncertainty.: To capture distributional shift, we define the extrapolation uncertainty as

$$\epsilon(q) = 1 - \max_{x_i \in D_{\text{train}}} \frac{x^\top x_i}{\|x\|_2 \, \|x_i\|_2}, \quad (5)$$

which measures the distance of q from the router's calibration manifold. This quantity can be efficiently approximated using nearest-neighbor indexing or prototype representations.

The above signals are synthesized into a scalar risk proxy

$$\tilde{r}(q) = w(q) \cdot \sigma\big(\gamma_0 + \gamma_1\big(1 - p(q)\big) + \gamma_2 u(q) + \gamma_3 \epsilon(q)\big)\,, \quad (6)$$

which serves as a compact abstraction of the query's latent exposure to catastrophic failure. The parameters $\{\gamma_i\}$ are learned on a small calibration set using logistic regression or an equivalent convex objective, incurring negligible tuning overhead. Here, $w(q)$ encodes failure severity: it defines the true loss magnitude in $Z(q,l)$ for tail-risk characterization, and modulates decision-time risk intensity in $\tilde{r}(q)$ to prioritize high-stakes queries during routing. We view $\tilde{r}(q)$ as a query-level proxy for *tail membership*, i.e., a monotone score correlated with $\Pr(Z(q,1) \geq \tau)$ for high thresholds τ, as estimated from calibration data. Consequently, minimizing a surrogate objective that upweights queries with larger $\tilde{r}(q)$ emphasizes the upper quantiles of the loss distribution, thereby aligning the induced routing behavior with CVaR-style tail-risk control under the one-shot decision constraint.

D. Adaptive Risk-Aware Routing Policy

Since CVaR is dominated by extreme realizations of $Z(q,l)$, effective tail-risk control requires prioritizing queries with high latent risk exposure. Direct optimization of CVaR is infeasible under the one-shot constraint. We therefore minimize a calibrated surrogate that upper-bounds tail risk by upweighting high-risk queries.

Specifically, we define the joint decision score

$$S(q,l) = \lambda\, c(l) + \beta\, \tilde{r}(q)\, \rho(l), \tag{7}$$

where λ controls the global inference budget and β encodes the system's risk-aversion level. This formulation decouples economic constraints from safety preferences, allowing the router to prioritize reliability on high-stakes queries even under tight cost limits. The term $\rho(l)$ is a *knowledge-level structural prior* that captures the attenuation of tail risk afforded by higher-capability models. We emphasize that the parameters $\{\gamma_i\}$ in the risk proxy are learned during a lightweight calibration phase and remain fixed at inference time, whereas λ and β are user-specified deployment-time hyperparameters that regulate the cost–risk trade-off.

We adopt the monotonic form

$$\rho(l) = \exp(-\kappa(l-1))\,, \tag{8}$$

where $\kappa > 0$ is a decay constant controlling the rate at which tail risk is attenuated across model tiers. This form encodes the prior belief that higher-capability models reduce the magnitude of tail losses. This monotone decay is empirically consistent with the common observation that higher-tier models exhibit systematically lower failure rates and reduced loss variance under distributional shift. In practice, the decay parameter κ can be calibrated on a held-out validation set by fitting the observed tail-risk–cost trade-off across model tiers, while enforcing monotonicity to preserve interpretability and structural consistency. While chosen for clarity, $\rho(l)$ can be learned or calibrated from validation data, provided that monotonicity across the model hierarchy is preserved. The routing policy then selects the tier minimizing the joint decision score:

$$R(q) = \arg \min_{l \in \{1,\dots,L\}} S(q,l). \tag{9}$$

Defensive Reasoning Principle. As predictive or extrapolation uncertainty increases, the reliability of performance prediction degrades, causing the risk surrogate $\tilde{r}(q)$ to dominate the decision score and bias routing toward higher-capability models despite higher cost. This induces an implicit safety envelope that prioritizes robustness where epistemic knowledge is weakest. Under mild assumptions that $\tilde{r}(q)$ is rank-consistent with tail events of $Z(q,l)$, minimizing the surrogate score yields effective control of global CVaR while respecting inference budget constraints.

IV. Experiments

We conduct extensive experiments to evaluate the proposed risk-aware one-shot routing framework. Routing is treated as an instance-level decision process balancing response quality, inference cost, and tail-risk reliability, with a particular focus on out-of-distribution (OOD) and safety-critical settings. All methods are evaluated under a unified three-tier model hierarchy and a strict one-shot constraint. We investigate the following research questions (RQs):

RQ1: How does the proposed router shape the quality–cost Pareto frontier compared to state-of-the-art multi-tier routing methods?

RQ2: Does the decision-theoretic CVaR objective effectively suppress catastrophic failures in high-risk and safety-critical domains?

RQ3: How do performance prediction, predictive uncertainty, and extrapolation risk jointly influence routing decisions under epistemic uncertainty?

RQ4: How do routing behaviors adapt to varying query risk intensities relative to fixed-threshold or heuristic strategies?

RQ5: How stable and controllable is the framework under different cost–risk priority configurations?

We further include a qualitative case study to illustrate how multi-dimensional risk signals are integrated for context-aware model selection in real-world scenarios.

A. Experimental Setup

Evaluation Benchmarks and Risk Semantics. We evaluate the proposed framework on benchmarks spanning general reasoning, professional decision-making, and safety-critical scenarios. Reasoning tasks include GSM8K [23] and MATH [24], while general knowledge is assessed on MMLU [25] and SimpleQA [26]. Professional domains are represented by MedQA [27] and FinQA [28], and safety robustness is evaluated on WildChat [29] and JailbreakBench [30]. The FLAN-v2 [31] instruction mixture is used for router training and calibration. All evaluation benchmarks are treated as out-of-distribution (OOD) relative to the training data.

Model Action Space. We construct a three-tier hierarchical model space reflecting cost–reliability trade-offs. Tier-1 (Lightweight) includes Llama-3-8B [32], Qwen2-7B [33], and Gemma-2-9B [34]. Tier-2 (Mid-range) consists of Llama-3-70B, GPT-4o-mini [35], and Mixtral-8x22B [36], while Tier-3 (Frontier) includes GPT-4o, Claude-3.5-Sonnet, and DeepSeek-R1 [37]. Routing decisions are made at the tier level using a fixed representative model per tier to ensure a consistent capability hierarchy.

Baselines and Evaluation Protocol. We compare our approach against three routing paradigms: (i) uncertainty-aware methods relying on confidence or uncertainty estimation (RouterDC [17], CP-Router [18], CARGO [19]); (ii) optimization-based methods balancing inference cost and expected quality (RouteLLM [2], FrugalGPT [14]); and (iii) heuristic or cascade-based strategies (HybridLLM [3], AutoMix [9]). All baselines are adapted to the unified model pool and evaluated under a strict one-shot constraint. We adopt BGE-M3 [38] as a frozen encoder to estimate extrapolation risk via maximum cosine similarity to the calibration set. Predictive uncertainty is estimated using stochastic forward passes. Failure severity weights $w(q)$ encode prior domain knowledge, with higher

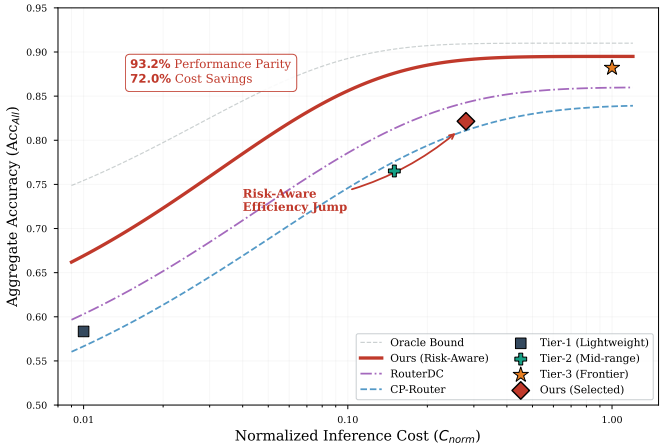

Fig. 3. Quality–cost Pareto frontier under three-tier routing. Risk-aware routing achieves a pronounced efficiency gain in the intermediate cost regime, approaching frontier performance at substantially reduced inference cost.

values assigned to professional and safety-critical queries. Evaluation metrics include Accuracy–Cost Pareto curves and CVaR@95%, which directly measures tail-risk reliability. All inference costs are normalized relative to Tier-1 pricing.

B. RQ1: Overall Performance and Cost-Efficiency Analysis

We evaluate the proposed risk-aware one-shot router across reasoning, professional, and safety-critical benchmarks. As shown in Fig. 3, our method consistently forms the upper envelope of the quality–cost Pareto frontier over a wide range of inference budgets. The most pronounced gains occur in the intermediate cost regime around $C_{\text{norm}} \approx 0.28$, where the router selectively escalates high-stakes queries and bridges the gap between lightweight and frontier models. In this regime, our approach recovers 93.2% of Tier-3 performance while using only 28.0% of the inference cost, yielding a clear efficiency jump. In contrast, optimization-based and uncertainty-aware baselines such as RouteLLM, RouterDC, and CARGO exhibit degraded performance at comparable cost levels, reflecting their limited ability to account for heterogeneous failure severity at the instance level. Beyond average accuracy, Table I shows substantial improvements in system-level reliability. Compared with the strongest uncertainty- and confidence-aware baselines, our method reduces CVaR@95% from 0.524 (RouterDC) and 0.503 (CARGO) to 0.245. On high-risk queries, it achieves 81.40% accuracy, approaching Tier-3 performance (86.40%) at less than one-third of the inference cost. Robustness under distributional shift is further evidenced by the OOD failure rate, which is reduced to 6.20%, substantially outperforming cascade-based and uncertainty-aware baselines. These results demonstrate that explicitly modeling extrapolation uncertainty and failure severity is critical for achieving reliable and cost-efficient LLM routing in safety-critical settings.

TABLE I
OVERALL PERFORMANCE, COST EFFICIENCY, AND RELIABILITY AVERAGED ACROSS ALL BENCHMARKS

Category	Method	Acc_{All} ↑	Acc_{High} ↑	CVaR↓	C_{norm} ↓	OOD Fail↓
Single Tier	Tier-1	58.40	42.50	0.842	0.010	44.50
	Tier-2	76.50	68.20	0.485	0.150	18.20
	Tier-3	88.20	86.40	0.125	1.000	1.80
Baselines	FrugalGPT	72.40	62.10	0.612	0.250	22.10
	AutoMix	74.80	64.50	0.584	0.300	19.40
	CP-Router	77.20	70.40	0.542	0.320	14.20
	RouteLLM	78.50	72.80	0.518	0.350	15.60
	RouterDC	79.20	74.50	0.524	0.380	12.80
	CARGO	79.80	74.98	0.503	0.369	11.70
Ours	**Risk-Aware**	**82.20**	**81.40**	**0.245**	**0.280**	**6.20**

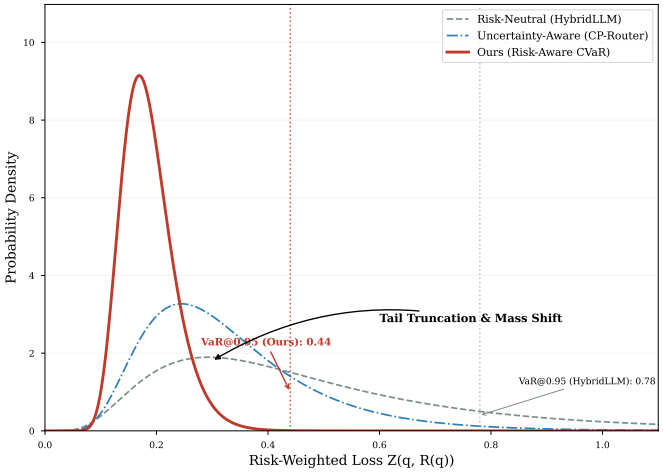

Fig. 4. Probability density of the risk-weighted loss $Z(q, R(q))$. Risk-aware routing truncates the heavy tail, reducing high-impact failures compared to expectation- and uncertainty-based baselines.

C. RQ2: Tail-Risk Mitigation and Reliability Analysis

We assess the ability of the proposed framework to suppress catastrophic failures by analyzing the distribution of the risk-weighted loss $Z(q, R(q))$ on high-stakes benchmarks (MedQA, FinQA, and JailbreakBench), where elevated query weights encode severe failure consequences. Figure 4 compares the resulting loss distributions across routing strategies. As shown in Fig. 4, expectation-based routing (HybridLLM) produces a pronounced heavy tail, indicating frequent exposure to rare but high-impact failures. Uncertainty-aware routing (CP-Router) achieves only limited tail reduction, as confidence signals alone are insufficient to capture heterogeneous failure severity. In contrast, the proposed risk-aware router explicitly prioritizes tail-risk control, shifting probability mass toward the low-loss region and sharply truncating the tail. As a result, the 95th-percentile Value-at-Risk is reduced from 0.78 (HybridLLM) and 0.62 (CP-Router) to 0.44. This substantial truncation of the loss distribution tail demonstrates that integrating predictive uncertainty, extrapolation risk, and failure severity within a CVaR-oriented routing objective is critical for bounding worst-case risk, particularly in safety-critical and distribution-shifted settings.

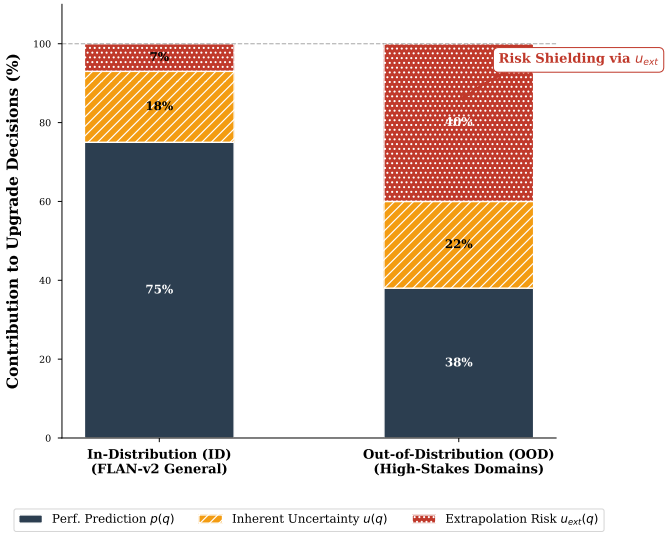

Fig. 5. Decomposition of risk proxy contributions under in-distribution and out-of-distribution settings. Extrapolation risk dominates under distributional shift, compensating for degraded performance-based signals.

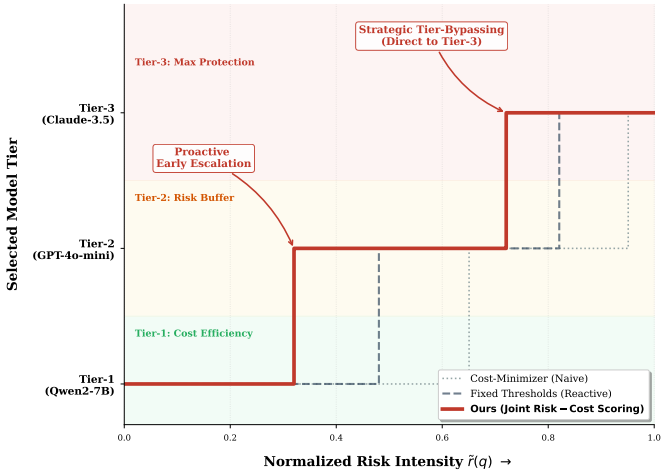

Fig. 6. Adaptive decision trajectories under varying risk intensity in a three-tier hierarchy. Joint risk–cost scoring enables proactive escalation and non-linear tier bypassing compared to threshold-based baselines.

D. RQ3: Decomposing the Contribution of Multi-dimensional Risk Proxies

We analyze the contribution of individual risk proxies—performance prediction $p(q)$, predictive uncertainty $u(q)$, and extrapolation risk $\epsilon(q)$—to routing decisions under in-distribution (FLAN-v2) and out-of-distribution (OOD) settings. Figure 5 summarizes the decomposition results. Under in-distribution conditions, routing decisions are dominated by performance prediction, which accounts for approximately 75% of model escalations, indicating that accuracy-based predictors remain reliable within the training manifold. Under distributional shift, however, the contribution of performance prediction drops sharply to 38%. In contrast, extrapolation risk $\epsilon(q)$ becomes the primary driver, increasing from 7% in-distribution to 40% under OOD conditions, thereby compensating for the degradation of performance-based signals. Predictive uncertainty $u(q)$ serves as a stable auxiliary cue across both regimes, contributing 18–22% but remaining insufficient on its own to identify high-risk queries. These results confirm that effective risk-aware routing requires integrating complementary epistemic signals, and that extrapolation risk is essential for preserving risk sensitivity when performance predictors become unreliable, consistent with the tail-risk mitigation observed in RQ2.

E. RQ4: Adaptive Decision Trajectories under Varying Risk Intensities

We analyze routing dynamics by visualizing decision trajectories as a function of normalized risk intensity. Figure 6 compares the proposed joint risk–cost scoring policy with cost-only and fixed-threshold baselines under a three-tier hierarchy. As shown in Fig. 6, the proposed policy exhibits proactive and non-linear escalation behavior. Queries are upgraded from Tier-1 to Tier-2 once the risk intensity reaches 0.32, substantially earlier than fixed-threshold routing, which remains at Tier-1 until risk exceeds 0.48. As risk increases further, the policy demonstrates strategic tier bypassing: when risk exceeds 0.72, queries are frequently routed directly to Tier-3, skipping Tier-2. This leapfrogging behavior arises from the exponential risk modulation in the joint decision score, which prioritizes reliability once latent risk crosses a critical level. Unlike linear cascade strategies, the proposed approach operates under a strict one-shot constraint, enabling timely and context-aware model selection without cumulative inference latency. These results confirm that joint risk–cost scoring supports both efficient resource utilization and robust protection under rapidly varying risk conditions.

F. RQ5: Parameter Sensitivity and Operational Stability

We evaluate the sensitivity of the proposed framework to the risk sensitivity coefficient β and the cost weight λ using a mixed workload spanning FLAN-v2 and MedQA. Figure 7 shows that the normalized tail risk varies smoothly over the joint parameter space. As illustrated, increasing β consistently strengthens tail-risk suppression by amplifying risk signals and favoring higher-reliability models, whereas increasing λ enforces stricter cost control and leads to a gradual rise in tail risk as lower-cost tiers are preferred. Importantly, these effects are continuous and monotonic, indicating stable and predictable control of the cost–risk trade-off. The heatmap reveals a broad operational region (approximately $\beta \in [2.5, 3.5]$ and $\lambda \in [0.1, 0.4]$) in which the framework achieves a favorable balance between inference cost and tail-risk mitigation without abrupt performance degradation. This smooth behavior contrasts with fixed-threshold heuristics that are highly sensitive to precise tuning, and demonstrates that the proposed router can be reliably calibrated across diverse deployment requirements without exhaustive hyperparameter search.

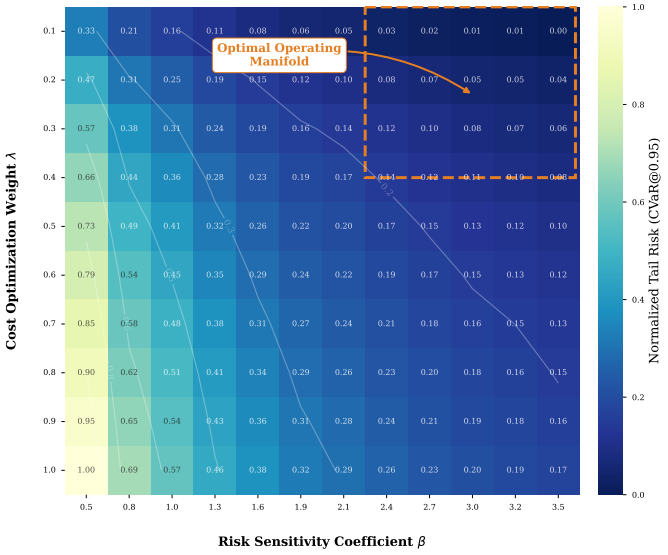

Fig. 7. Sensitivity of normalized tail risk to the risk sensitivity coefficient β and cost weight λ. Smooth variation reveals a broad operational region balancing cost efficiency and risk mitigation.

TABLE II
QUALITATIVE CASE STUDIES SHOWING HOW CONSTITUENT RISK SIGNALS INFLUENCE ROUTING DECISIONS

Domain	Query Example	$1-p(q)$	$u(q)$	$\epsilon(q)$	$w(q)$	Selected Tier
Medical	Pediatric dosage for rare condition	Med.	Med.	High	1.5	Tier-3
General	Capital of historical province	Med.	Med.	Low	0.5	Tier-1
Finance	Volatile derivative ROI analysis	Med.	Med.	Med.	1.2	Tier-3
Safety	Dual-use chemical synthesis query	High	High	High	2.0	Tier-3

G. Qualitative Analysis: Case Studies in Risk-Aware Reasoning

To illustrate instance-level routing behavior, Table II reports representative case studies together with the constituent epistemic risk signals—upgrade necessity $1-p(q)$, predictive uncertainty $u(q)$, extrapolation risk $\epsilon(q)$, and failure severity weight $w(q)$—that are integrated into the unified risk proxy $\tilde{r}(q)$. The examples highlight three characteristic decision patterns. In a high-stakes medical query, elevated extrapolation risk and failure severity dominate despite moderate performance predictions, leading to direct escalation to Tier-3 for safety. In contrast, a low-severity general knowledge query is routed to Tier-1 even under non-negligible uncertainty, prioritizing cost efficiency. A financial analysis query illustrates strategic tier bypassing, where moderate extrapolation risk combined with increased severity triggers direct selection of Tier-3 over intermediate models. Across cases, routing decisions do not rely on fixed accuracy or uncertainty thresholds, but emerge from the structured integration of multiple risk signals, consistent with the decision-theoretic formulation in Section III-C.

V. Conclusion

This work formulates large language model routing as an instance-level, risk-aware decision-making problem and demonstrates that optimizing expected performance alone is insufficient when failure consequences are heterogeneous. By casting routing as a cost-aware, decision-theoretic optimization task, we explicitly target tail-risk suppression under constrained inference budgets. We introduce lightweight and interpretable risk abstractions that enable a practical one-shot routing policy, balancing reliability and cost without static thresholds or multi-model cascading. Extensive experiments show that the proposed framework substantially improves reliability on high-stakes and out-of-distribution queries while maintaining competitive efficiency. These results highlight the importance of explicit risk-aware reasoning for robust and explainable LLM deployment. Future work will investigate richer epistemic risk representations and extensions to non-stationary and adaptive deployment environments.

Acknowledgment

The authors acknowledge the use of AI-assisted tools for language editing and clarity improvement during the preparation of this manuscript. The authors have reviewed and verified all content and take full responsibility for the accuracy and integrity of the work.

Funding

This work is supported by National Natural Science Foundation of China (Grant Nos. 62371069, 62272056, 62372048).

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, and I. Stoica, "Routellm: Learning to route llms with preference data," *arXiv preprint arXiv:2406.18665*, 2024.

[3] D. Ding, A. Mallick, C. Wang, R. Sim, S. Mukherjee, V. Ruhle, L. V. Lakshmanan, and A. H. Awadallah, "Hybrid llm: Cost-efficient and quality-aware query routing," *arXiv preprint arXiv:2404.14618*, 2024.

[4] Y. Mei, Y. Zhuang, X. Miao, J. Yang, Z. Jia, and R. Vinayak, "Helix: Serving large language models over heterogeneous gpus and network via max-flow," in *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, 2025, pp. 586–602.

[5] K. Wang, M. Guo, C. Dai, and Z. Li, "Information-decision searching algorithm: Theory and applications for solving engineering optimization problems," *Information Sciences*, vol. 607, pp. 1465–1531, 2022.

[6] R. T. Rockafellar, S. Uryasev *et al.*, "Optimization of conditional value-at-risk," *Journal of risk*, vol. 2, pp. 21–42, 2000.

[7] C. Y.-C. Chen, J. Shen, Z. Deng, and L. Lei, "Conformal tail risk control for large language model alignment," *arXiv preprint arXiv:2502.20285*, 2025.

[8] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," *Advances in neural information processing systems*, vol. 32, 2019.

[9] P. Aggarwal, A. Madaan, A. Anand, S. P. Potharaju, S. Mishra, P. Zhou, A. Gupta, D. Rajagopal, K. Kappaganthu, Y. Yang *et al.*, "Automix: Automatically mixing language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 131 000–131 034, 2024.

[10] X. Wang, Y. Liu, W. Cheng, X. Zhao, Z. Chen, W. Yu, Y. Fu, and H. Chen, "Mixllm: Dynamic routing in mixed large language models," *arXiv preprint arXiv:2502.18482*, 2025.

[11] A. Tsiourvas, W. Sun, and G. Perakis, "Causal llm routing: End-to-end regret minimization from observational data," *arXiv preprint arXiv:2505.16037*, 2025.

[12] S. Lee, D. B. Lee, D. Wagner, M. Kang, H. Seong, T. Bocklet, J. Lee, and S. J. Hwang, "Saferoute: Adaptive model selection for efficient and accurate safety guardrails in large language models," *arXiv preprint arXiv:2502.12464*, 2025.

[13] S. Li, L. Wu, T. Peng, J. Huang, H. Jiang, Y. Xue, J. Zhang, and D. W. Gao, "Rsynllm: A risk-aware routing mixture-of-experts large language model for multi-energy load forecasting in large-scale distribution networks," *Applied Energy*, vol. 406, p. 127227, 2026.

[14] L. Chen, M. Zaharia, and J. Zou, "Frugalgpt: How to use large language models while reducing cost and improving performance," *arXiv preprint arXiv:2305.05176*, 2023.

[15] D. Soiffer, S. Kolawole, and V. Smith, "Semantic agreement enables efficient open-ended llm cascades," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2025, pp. 2499–2537.

[16] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson *et al.*, "Language models (mostly) know what they know," *arXiv preprint arXiv:2207.05221*, 2022.

[17] S. Chen, W. Jiang, B. Lin, J. Kwok, and Y. Zhang, "Routerdc: Query-based router by dual contrastive learning for assembling large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 66 305–66 328, 2024.

[18] J. Su, F. Lin, Z. Feng, H. Zheng, T. Wang, Z. Xiao, X. Zhao, Z. Liu, L. Cheng, and H. Wang, "Cp-router: An uncertainty-aware router between llm and lrm," *arXiv preprint arXiv:2505.19970*, 2025.

[19] A. Barrak, Y. Fourati, M. Olchawa, E. Ksontini, and K. Zoghlami, "Cargo: A framework for confidence-aware routing of large language models," *arXiv preprint arXiv:2509.14899*, 2025.

[20] J. Y. Halpern, *Reasoning about uncertainty*. MIT press, 2017.

[21] W. Overman and M. Bayati, "Conformal arbitrage: Risk-controlled balancing of competing objectives in language models," *arXiv preprint arXiv:2506.00911*, 2025.

[22] S. Pan and D. Wu, "Trustworthy summarization via uncertainty quantification and risk awareness in large language models," *arXiv preprint arXiv:2510.01231*, 2025.

[23] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[24] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset," *arXiv preprint arXiv:2103.03874*, 2021.

[25] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[26] J. Wei, N. Karina, H. W. Chung, Y. J. Jiao, S. Papay, A. Glaese, J. Schulman, and W. Fedus, "Measuring short-form factuality in large language models," *arXiv preprint arXiv:2411.04368*, 2024.

[27] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.

[28] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. R. Routledge *et al.*, "Finqa: A dataset of numerical reasoning over financial data," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3697–3711.

[29] W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng, "Wildchat: 1m chatgpt interaction logs in the wild," *arXiv preprint arXiv:2405.01470*, 2024.

[30] P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Sehwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramer *et al.*, "Jailbreakbench: An open robustness benchmark for jailbreaking large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 55 005–55 029, 2024.

[31] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

[32] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[33] Q. Team *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, vol. 2, no. 3, 2024.

[34] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.

[35] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[36] Y. Bendi-Ouis, D. Dutartre, and X. Hinaut, "Deploying open-source large language models: A performance analysis," *arXiv preprint arXiv:2409.14887*, 2024.

[37] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[38] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," *arXiv preprint arXiv:2402.03216*, vol. 4, no. 5, 2024.