Error-Aware Reverse Auction Mechanism for Large Language Model Routing

Haolong Chen[1,2,3,*], **Zhengyuan Xin**[2,3,*], **Liang Zhang**[4], **Lei Xue**[4], **Guangxu Zhu**[1,2,3,5]
[1]Shenzhen International Center for Industrial and Applied Mathematics
[2]Shenzhen Research Institute of Big Data
[3]The Chinese University of Hong Kong, Shenzhen
[4]Shenzhen Campus of Sun Yat-sen University
[5]Shenzhen Loop Area Institute
[*]Equal Contribution
{haolongchen1, zhengyuanxin}@link.cuhk.edu.cn,
{zhangliang27, xuelei3}@mail.sysu.edu.cn, gxzhu@sribd.cn

Abstract

Routing each query to a cost-effective large language model (LLM) is critical for balancing quality and cost, yet most routers rely on a centralized task center to predict model performance, creating an information-risk mismatch and a scalability bottleneck as the model pool grows. We propose a market-based routing paradigm that shifts ex-ante prediction to LLM providers via a reverse auction, where providers bid with self-predicted success probabilities and execution costs. To account for inherently noisy provider predictions and center evaluations, we introduce the ***E**rror-**A**ware **R**everse **A**uction **M**echanism* (EA-RAM), which explicitly models this inherent Dual Error. We prove that EA-RAM is Bayesian incentive compatible and individually rational under the Dual Error, establish sufficient conditions for center rationality, and derive an explicit welfare-loss bound. We further identify robustness effects: opposite-signed errors can cancel, vanishing-tail link functions (e.g., logistic) stabilize clear-cut cases via saturation, and extra noise smooths belief maps, reducing the gains from marginal manipulation. Experiments on simulations and real-world benchmarks show that EA-RAM is robust to the Dual Error and achieves a better cost–performance Pareto frontier than centralized baselines, with additional gains when providers contribute local information, validating its practical effectiveness.

1 Introduction

Large language models (LLMs) underpin many intelligent applications [1, 2, 3], yet the expanding model ecosystem makes it increasingly necessary to route each query to a cost-effective model. Because model capabilities and costs vary widely, from expensive frontier models to cheaper specialized ones, *LLM routing* is essential for optimizing the cost–performance trade-off.

Most existing routers follow a centralized estimation paradigm,

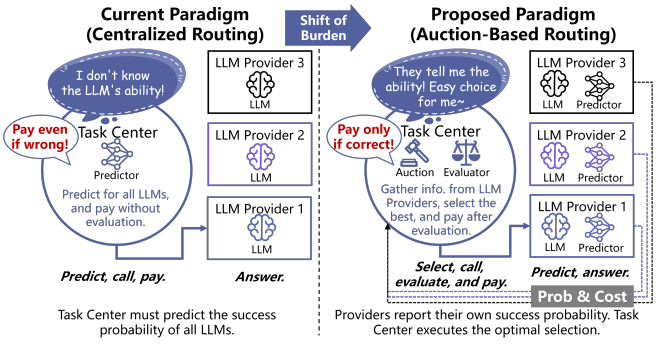


Figure 1: Paradigm shift from traditional Centralized Routing to our proposed Auction-Based Routing.

Preprint.

where the task center predicts each model's performance [4, 5, 6]. It has two structural limitations. *(i) Information–risk mismatch:* the task center bears the risk of failure but has less information about the LLMs, whereas providers hold richer private knowledge. Commercial routers[1] largely retain this paradigm and thus inherit the same inefficiency. *(ii) Scalability bottleneck:* the center must profile or train for every new model, incurring per-model overhead even for training-free routers [6], which hinders rapid expansion.

To address these limitations, we propose a market-based routing paradigm that transfers ex-ante prediction to LLM providers via a reverse auction (Figure 1). The task center acts as the buyer and solicits bids from providers, who report their self-predicted success probabilities and costs; the center maintains only a model-agnostic ex-post evaluator. This distributed design aligns information with risk and removes the need for model-by-model profiling for the buyer, thereby improving scalability.

Similar paradigm shifts have proven successful in other mature domains. For example, moving from static allocation to market-based mechanisms such as real-time bidding in advertising [9, 10] and auction-based spectrum allocation [11, 12]. However, transferring this idea to LLM routing is *not* straightforward. Designing such a mechanism requires a new theory because LLM routing operates under an inherent *Dual Error*: providers' ex-ante success estimates are subjective and noisy, and the center's ex-post evaluation is imperfect. This setting departs from prior fault-tolerant allocation mechanisms [13, 14, 15], which typically assume *ideal observability*—i.e., task outcomes are objectively verifiable and/or agents' success probabilities are common knowledge—and thus can distort incentives if such assumptions are violated.

We therefore propose the ***E**rror-**A**ware **R**everse **A**uction **M**echanism* (EA-RAM) Crucially, EA-RAM explicitly models the inherent Dual Error by characterizing both providers' subjective error in ex-ante prediction and the task center's error in ex-post evaluation. We characterize equilibrium bidding and prove that EA-RAM is Bayesian incentive compatible (BIC) and individually rational (IR) under Dual Error; we further provide sufficient conditions for center rationality (CR) and derive an explicit upper bound on welfare loss relative to the error-free benchmark. Beyond these, we uncover three structural robustness effects: ex-ante and ex-post errors with opposite signs can offset each other; when the link function has vanishing tails (e.g., logistic), large-margin (clear-cut) instances lie in the saturated region and are thus insensitive to noise; and extra independent noise smooths the belief maps, reducing their maximum slope and thereby limiting the gains from marginal manipulation. Extensive simulations and real-world experiments show that EA-RAM is robust to Dual Error and achieves a superior cost–performance trade-off over state-of-the-art centralized baselines, especially when leveraging providers' local information. Our main contributions are summarized as follows:

1. We propose EA-RAM, a market-based routing framework that shifts ex-ante prediction to LLM providers, addresses centralized routers' information–risk mismatch and per-model profiling bottleneck, and models LLM routing as an auction under explicit Dual Error.
2. We establish theoretical guarantees under Dual Error: EA-RAM satisfies BIC and IR, admits sufficient conditions for CR, and enjoys a welfare-loss bound; we further reveal robustness insights, including error compensation, saturation stability, and noise-induced flattening.
3. Extensive experiments on simulations and real-world benchmarks show that EA-RAM is robust to Dual Error and improves economic efficiency compared to state-of-the-art centralized baselines.

2 Error-Aware Reverse Auction Mechanism

We formulate LLM routing as a strategic reverse auction between the task center and LLM providers. As shown in Figure 2, providers submit ex-ante bids, the center allocates the query, and payments are settled after ex-post evaluation.

2.1 Basic Setting

We consider a task center (the *buyer*) and a set of risk-neutral[2] LLM providers (the *sellers*) $\mathcal{I} = \{1, \ldots, N\}$ over independent, non-combinatorial *tasks* $t \in \mathcal{T}$. Hence, we omit the superscript t. Each

[1] e.g., OpenRouter [7] and Requesty [8]

[2] Risk neutrality is reasonable in repeated, high-volume platform settings, where providers optimize long-run expected profit and single-query risk is small relative to their portfolio.

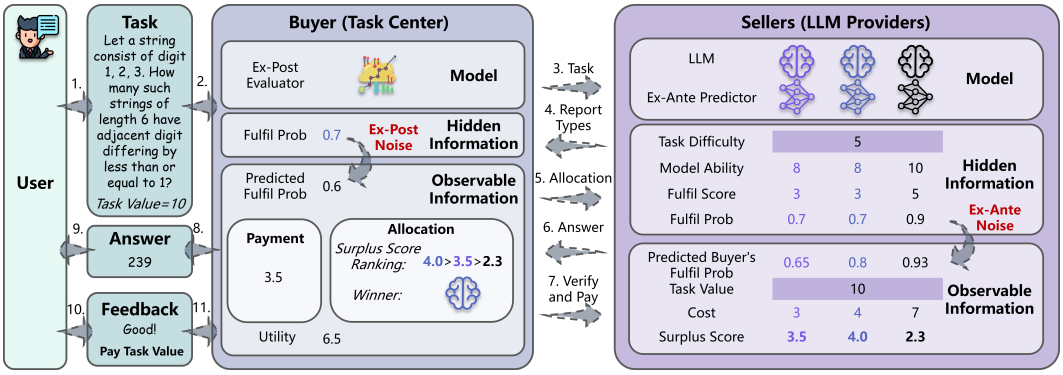


Figure 2: EA-RAM for LLM routing. (1) Providers bid using ex-ante predictions. (2) The buyer allocates the query by reported surplus. (3) The selected model answers. (4) The buyer evaluates the output and settles payment. (5) The user receives accepted output and pays the task value.

task has a common-knowledge *value* $V > 0$, realized by the buyer iff the task demand is fulfilled ($\mu = 1$). Seller i has private *type* $\theta_i = (p_i, c_i)$, where $c_i > 0$ is the *execution cost* and p_i is the true *fulfillment probability* of the binary *fulfillment indicator* $\mu_i \sim \text{Bernoulli}(p_i)$.

Following prior work on capability modeling [6, 16], we model the *fulfillment probability* p_i via the alignment between seller i's *model ability* $m_i \in \mathbb{R}$ and the *task difficulty* $d \in \mathbb{R}$ (Figure 3). The latent *fulfillment score* $\phi_i = \phi(m_i, d)$ satisfies $\partial\phi/\partial m_i \geq 0$ and $\partial\phi/\partial d \leq 0$, and $p_i = \sigma(\phi_i)$, where the *link function* $\sigma : \mathbb{R} \to (0, 1)$ is strictly increasing, continuously differentiable, and globally L_σ-Lipschitz. Neither p_i nor μ_i is observed during mechanism execution.

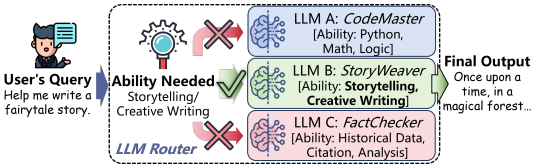


Figure 3: Ability–Difficulty Matching.

2.2 Error-Aware Prediction and Evaluation

LLM Routing faces *Dual Error* due to noisy prediction and imperfect evaluation, undermining classical mechanisms such as FTMD [13] and necessitating an error-aware formulation.

Buyer's ex-post evaluation. The buyer evaluates the output via the error-involved *ex-post acceptance probability* $h_i = \sigma(\phi_i + \varepsilon_{\text{post}})$, where the *evaluation error* $\varepsilon_{\text{post}}$ has $\mathbf{E}[\varepsilon_{\text{post}}] = a_{\text{post}}$ and $\text{Var}(\varepsilon_{\text{post}}) = b_{\text{post}}$. The decision $\tilde{\mu}_i \sim \text{Bernoulli}(h_i)$ determines whether the answer is sent to the user.

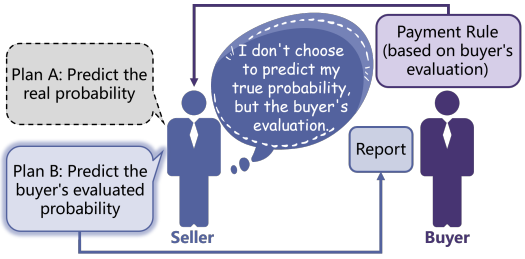


Figure 4: Seller's strategic decision. The seller bids based on the buyer's error-involved evaluation, rather than the unobserved ground-truth.

Sellers' ex-ante prediction. Since payments depend on the buyer's evaluation, seller i predicts the buyer's acceptance probability rather than the ground-truth fulfillment probability. We introduce an independent *prediction error* $\varepsilon_{\text{ante},i}$ with $\mathbf{E}[\varepsilon_{\text{ante},i}] = a_{\text{ante},i}$ and $\text{Var}(\varepsilon_{\text{ante},i}) = b_{\text{ante},i}$, and define the *aggregated error* $\eta_i = \varepsilon_{\text{post}} + \varepsilon_{\text{ante},i}$. Thus the seller's *ex-ante subjective probability* is $g_i = \sigma(\phi_i + \eta_i)$, where $\mathbf{E}[\eta_i] = a_{\eta,i} = a_{\text{post}} + a_{\text{ante},i}$ and $\text{Var}(\eta_i) = b_{\eta,i} = b_{\text{post}} + b_{\text{ante},i}$. As shown in Figure 4, sellers bid based on g_i rather than p_i, a rationale formalized in Section 3.2.

2.3 Mechanism Interaction

A *mechanism* is a mapping $\Gamma : \hat{\theta} \mapsto (f(\hat{\theta}), r(\hat{\theta}, \tilde{\mu}))$, where f is the allocation rule and r is the payment rule, based on sellers' reports $\hat{\theta}$ and the evaluator signal $\tilde{\mu}$. Each seller submits a report $\hat{\theta}$; the buyer applies $f(\hat{\theta})$ to select a winner or the null allocation; after execution, the buyer observes

$\tilde{\mu}$ and settles transfers via $r(\hat{\theta}, \tilde{\mu})$. Utilities and social welfare are then computed from the realized allocation, evaluation, and payment, as summarized in Algorithm 1.

Utilities and welfare. Our mechanism is welfare-oriented but strategic-agent-aware: the buyer aims to maximize expected social welfare, whereas each seller aims to maximize its own expected utility under the induced allocation and payment rules. Any non-winning seller $i \neq j$ obtains zero utility. The winner j receives payment r_j, incurs cost c_j, and obtains $U_j^{\text{seller}} = r_j - c_j$, with $\mathbb{E}[U_j^{\text{seller}}] = \mathbb{E}[r_j] - c_j$. The buyer obtains $U^{\text{buyer}} = V\mu_j - r_j$, with $\mathbb{E}[U^{\text{buyer}}] = Vp_j - \mathbb{E}[r_j]$. Social welfare is $W = V\mu_j - c_j$ and $\mathbb{E}[W] = Vp_j - c_j$; equivalently, $W = U^{\text{buyer}} + \sum_{i=1}^{N} U_i^{\text{seller}}$ because transfers are internal.

Algorithm 1 EA-RAM

Input: Sellers $\{m_i, c_i\}_{i=1}^N$; task (d, V).
for $i = 1, \ldots, N$ **do**
  Compute (ϕ_i, p_i) and (h_i, g_i) with errors $\varepsilon_{\text{post}}$ and $\{\varepsilon_{\text{ante},i}\}_{i=1}^N$.
  Set ranking score $\hat{s}_i \leftarrow Vg_i - c_i$.
end for
$j \leftarrow \arg\max_i \hat{s}_i$; $H \leftarrow \max(0, \max_{k \neq j} \hat{s}_k)$.
if $\hat{s}_j \leq 0$ **then**
  return Null.
end if
Execute $\mu_j \sim \text{Bern}(p_j)$; evaluate $\tilde{\mu}_j \sim \text{Bern}(h_j)$.
Pay $r_j \leftarrow V\tilde{\mu}_j - H$; set $r_i \leftarrow 0$ for all $i \neq j$.
$U_j^{\text{seller}} \leftarrow r_j - c_j$; $U^{\text{buyer}} \leftarrow V\mu_j - r_j$.
$W \leftarrow V\mu_j - c_j$.
return j, $\{r_i\}_{i=1}^N$, U_j^{seller}, U^{buyer}, W.

Reports. Seller i privately observes cost c_i, forms an ex-ante belief g_i about the buyer's evaluation outcome, and reports a type $\hat{\theta}_i = (\hat{p}_i, \hat{c}_i)$. Since the allocation rule ranks sellers only through the induced *reported surplus score* $\hat{s}_i = V\hat{p}_i - \hat{c}_i$, any report can be equivalently summarized by $\hat{s}_i$; thus seller i's strategic choice reduces from the two-dimensional report $\hat{\theta}_i$ to a scalar $\hat{s}_i$. For incentive analysis under Dual Error, given belief g_i and true cost c_i, define the *error-adjusted effective surplus* $\bar{T}_i = Vg_i - c_i$, representing seller i's perceived ex-ante expected welfare contribution under the evaluation process. Theorem 3.2 will later show that, under our payment rule, it is optimal to align the reported score with this effective surplus, i.e., $\hat{s}_i = \bar{T}_i$.

Allocation. Given reports $\hat{\theta}$, the buyer allocates to maximize reported expected welfare. For seller i with report $\hat{\theta}_i = (\hat{p}_i, \hat{c}_i)$, the reported surplus is $V\hat{p}_i - \hat{c}_i = \hat{s}_i$; hence ranking by $\hat{s}_i$ is equivalent to ranking sellers by reported expected welfare. The rule selects the largest positive $\hat{s}_i$, or the null outcome if all reported surpluses are non-positive. Equivalently, introduce a dummy seller 0 with $(\hat{p}_0, \hat{c}_0) = (0, 0)$ and $\hat{s}_0 = 0$, representing the null allocation. Let $\bar{\mathcal{I}} = \mathcal{I} \cup \{0\}$ and $j \in \arg\max_{i \in \bar{\mathcal{I}}} \hat{s}_i$. Then $f(\hat{\theta}) = j$ if $j \neq 0$, and $f(\hat{\theta}) = \emptyset$ otherwise.

Payment and incentive alignment. Let j denote the winning seller and let $H = \max(0, \max_{k \neq j} \hat{s}_k)$ be the runner-up score. If payments depended on the self-reported probability $\hat{p}_i$, sellers could inflate $\hat{p}_i$ (e.g., report $\hat{p}_i = 1$). We therefore condition transfers on the evaluator signal $\tilde{\mu}_j \in \{0, 1\}$, a noisy proxy for the unobserved ground truth μ_j. The winner is paid $r_j = r(\hat{\theta}_j, \tilde{\mu}_j) = V\tilde{\mu}_j - H$ [3], while losers receive $r_i = 0$ for $i \neq j$. Here $V\tilde{\mu}_j$ rewards evaluated performance, and H charges the winner for the externality imposed on others, aligning the best response with surplus-based truthful competition (see Theorem 3.2 for the formal statement).

3 Mechanism Properties

This section establishes EA-RAM's theoretical foundations, showing that it remains incentive-aligned and economically robust under Dual Error. Proofs of all results are deferred to Appendix C.

3.1 Mechanism Goals

To ensure a reliable and sustainable routing market, we follow FTMD [13] and consider four standard desiderata: **(1) Dominant-Strategy Incentive Compatibility (DSIC):** Truth-telling maximizes each agent's expected utility regardless of others' actions. **(2) Individual Rationality (IR):** Truthful participation gives each seller non-negative expected utility. **(3) Center Rationality (CR):** The

[3] In practice, negative realized transfers can be handled via a prefunded reserve maintained by each participating seller. This implementation-layer safeguard does not change the mechanism analyzed below.

buyer's expected utility remains non-negative. **(4) Economic Efficiency (EE):** Under truthful reporting, the mechanism allocates to maximize total expected social welfare.

We formalize the *error-aware setting* as a unified framework covering two cases. The general noisy case is the *error-involved setting*; the ideal case with vanishing errors, $\varepsilon_{\text{post}} = 0$ and $\varepsilon_{\text{ante},i} = 0$, is the *error-free setting*, corresponding to FTMD [13]. In the error-free setting, ex-ante and ex-post beliefs coincide with the ground-truth probability, $g_i = h_i = p_i$, and the effective surplus report becomes the true surplus, $\hat{s}_i = \bar{T}_i = Vp_i - c_i$. We then establish the following benchmark property:

Proposition 3.1 (Optimality of the Error-Free Setting)**.** *In the error-free setting, the mechanism satisfies all four desirable properties: DSIC, IR, CR, and EE.*

Proposition 3.1 establishes the error-free benchmark, showing optimality under idealized conditions. Although Dual Error weakens these exact guarantees, **EA-RAM retains strong robustness in the error-involved setting: it satisfies BIC and IR, admits sufficient CR conditions when surplus margins absorb evaluation noise or evaluations are conservative, and bounds welfare loss**, ensuring controlled efficiency degradation under uncertainty.

3.2 Strategic Properties

Let H denote the runner-up score with cumulative distribution function F_H and probability density function f_H. For a unilateral report $\hat{s}_i$, the seller's interim expected utility is $U_i^{\text{seller}}(\hat{s}_i) = \Pr(H \le \hat{s}_i) \cdot \mathbf{E}[V\tilde{\mu}_i - H - c_i \mid H \le \hat{s}_i] = F_H(\hat{s}_i)(\bar{T}_i - \mathbf{E}[H \mid H \le \hat{s}_i])$, since $\mathbf{E}[V\tilde{\mu}_i \mid \cdot] = Vg_i$ from the seller's ex-ante perspective.

Theorem 3.2 (Bayesian Incentive Compatibility (BIC))**.** *Seller i's interim expected utility $U_i^{seller}(\hat{s}_i)$ is maximized at $\hat{s}_i = \bar{T}_i$. Hence, reporting $\hat{s}_i = \bar{T}_i$ is a Bayesian best response. Moreover, if $\bar{T}_i > 0$, then $U_i^{seller}(\hat{s}_i)$ is uniquely maximized at $\hat{s}_i = \bar{T}_i$.*

This implies that for every seller i, truthfully reporting $\hat{s}_i = \bar{T}_i$ is a Bayesian best response. Consequently, the strategy profile in which all sellers report $\hat{s}_i = \bar{T}_i$ constitutes a Bayesian Nash equilibrium. Under this report, seller i's expected utility is $\mathbf{E}[U_i^{\text{seller}}] = \int_{-\infty}^{\bar{T}_i} (\bar{T}_i - h)\, dF_H(h)$.

Theorem 3.3 (Individual Rationality (IR))**.** *At the equilibrium $\hat{s} = \bar{T}$, every seller satisfies IR: $\mathbf{E}[U_i^{seller} \mid \tilde{\theta}_i] \ge 0$.*

This implies that truthful reporting yields non-negative expected utility, thereby ensuring that participating in the auction remains a rational strategy for every seller.

Theorem 3.4 (Center Rationality (CR))**.** *Each of the following sufficient conditions guarantees CR: (A) $\mathbf{E}[H] \ge V\Delta_{gate}$, where $\Delta_{gate} = L_\sigma \sqrt{b_{post} + a_{post}^2}$; (B) $\Delta_{cons} = \mathbf{E}\big[p_{(1)} - h_{(1)}\big] \ge 0$.*

Intuitively, CR holds either (A) when the runner-up margin effectively buffers against evaluation noise, or (B) when the center's evaluation is sufficiently conservative. Under either condition, the mechanism is safe for the Task Center in the sense that its expected utility remains non-negative.

Proposition 3.5 (Comparative Statics: Ability and Difficulty)**.** *Holding fixed the runner-up score distribution F_H, seller i's interim expected utility $\mathbf{E}[U_i^{seller}]$ is weakly increasing in their model ability m_i and weakly decreasing in the task difficulty d.*

Intuitively, stronger models yield higher expected payoffs because they are more likely to meet the evaluator's standard, while increased task difficulty reduces potential gains.

3.3 Welfare Analysis

For welfare accounting, let the expected welfare under a specified selected seller i be $\mathbf{E}[W_i] = Vp_i - c_i$. We define the *true optimal winner index* (selected under the error-free setting) as $i^\star \in \arg\max_j \{Vp_i - c_i\}$ and the *error-aware setting's winner index* as $i^\dagger \in \arg\max_j \{Vg_j - c_j\}$.

Proposition 3.6 (Economic Efficiency (EE))**.** *The equilibrium allocation induced by the error-aware setting attains the same expected welfare as the error-free benchmark if and only if the error-aware winner $i^\dagger$ is welfare-optimal.*

Intuitively, Dual Error may distort the induced allocation away from the welfare-optimal seller, thereby causing welfare loss. We quantify this effect using the Dual Error's *second-moment radii*, $M_{\text{post}} = \sqrt{b_{\text{post}} + a_{\text{post}}^2}$ and $M_{\text{ante}} = \max_i \sqrt{b_{\text{ante},i} + a_{\text{ante},i}^2}$.

Theorem 3.7 (Welfare-loss bound)**.** *The expected welfare loss satisfies* $0 \le \mathbf{E}[W_{i^\star}] - \mathbf{E}[W_{i^\dagger}] = \mathbf{E}[(Vp_{i^\star} - c_{i^\star}) - (Vp_{i^\dagger} - c_{i^\dagger})] \le 2VL_\sigma(M_{\text{post}} + M_{\text{ante}})$.

This bound provides a guarantee of robustness, showing that the efficiency degradation scales linearly with the aggregate magnitude of the Dual Error and vanishes as the estimation quality improves.

3.4 Structural Insights

Having established EA-RAM's validity under Dual Error, we now examine its structural insights.

Proposition 3.8 (Opposite-Signed Error Compensation)**.** *If* $(g_i - h_i)(h_i - p_i) \le 0$*, for* $i \in \{i^\star, i^\dagger\}$*, then the welfare-loss bound tightens to* $2VL_\sigma \max\{M_{\text{ante}}, M_{\text{post}}\}$.

Intuitively, when ex-ante and ex-post errors have opposite signs, they partially cancel out, reducing the resulting welfare loss.

Proposition 3.9 (Saturation Robustness)**.** *In addition to the baseline link-function assumptions, suppose* σ *has vanishing tails, i.e.,* $\lim_{|x|\to\infty} \sigma'(x) = 0$*. For any error* ϵ *with* $\mathbf{E}[\epsilon^2] < \infty$*, the deviation* $\Delta(\phi_i) = |\mathbf{E}[\sigma(\phi_i + \epsilon)] - \sigma(\phi_i)|$ *satisfies* $\lim_{|\phi_i|\to\infty} \Delta(\phi_i) = 0$.

Intuitively, flat-tailed links such as the logistic function saturate when $|\phi_i|$ is large, so score noise barely affects $\sigma(\phi_i)$ and the mechanism is robust in clear-cut cases.

Proposition 3.10 (Noise-Induced Flattening)**.** *Let* η *be an independent noise term with* $\mathbf{E}[|\eta|] < \infty$*. Define the perturbed belief maps by convolution:* $\tilde{g}_i(\phi) = \mathbf{E}[g_i(\phi + \eta)]$ *and* $\tilde{h}_i(\phi) = \mathbf{E}[h_i(\phi + \eta)]$*. If* g_i *and* h_i *are continuously differentiable with bounded derivatives, then* $\sup_\phi |\frac{\partial \tilde{g}_i}{\partial \phi}(\phi)| \le \sup_\phi |\frac{\partial g_i}{\partial \phi}(\phi)|$ *and* $\sup_\phi |\frac{\partial \tilde{h}_i}{\partial \phi}(\phi)| \le \sup_\phi |\frac{\partial h_i}{\partial \phi}(\phi)|$*. Consequently, injecting additional independent noise weakly reduces the maximal sensitivity of both* $\phi \mapsto g_i(\phi)$ *and* $\phi \mapsto h_i(\phi)$.

Intuitively, ex-ante error and ex-post error lower the maximal sensitivity of the ranking map and the approval map, respectively, limiting the return to strategic manipulation.

4 Experiments

4.1 Simulation Experiments

In this section, we run simulation experiments that explicitly control the magnitude of the Dual Error to evaluate EA-RAM's robustness.

4.1.1 Experimental Setup

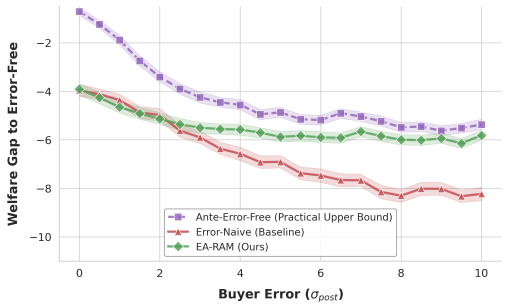
Seller	m_i	c_i	p_i	s_i
0	1.0	9.0	0.38	-1.45
1	1.5	10.0	0.50	0.00
2	2.5	12.0	0.73	2.62
3	3.5	12.8	0.88	4.81
4	4.5	13.8	0.95	5.25

Table 1: Simulation Settings.

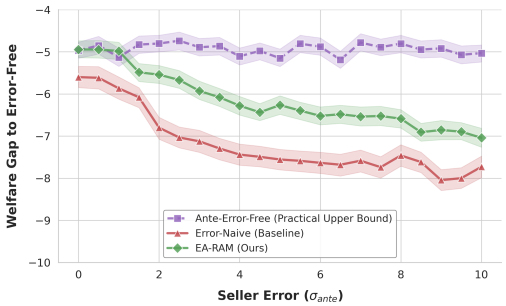
We simulate $N = 5$ heterogeneous sellers competing for one task with V=20 and d=1.5 (Table 1). To stress-test robustness, we use a highly competitive market with a small top-two surplus gap ($\Delta s = 0.44$), where mild noise can flip the ranking. Performance is measured by the *Welfare Gap* relative to the Error-Free upper bound, shown as the $y = 0$ line. We compare: (1) **Ante-Error-Free (Practical Upper Bound):** an oracle seller that perfectly models h_i and best responds by reporting $p_i + \epsilon_{post}$; (2) **Error-Naive (Baseline):** naive sellers that ignore evaluation and report $p_i + \epsilon_{ante}$; and (3) **EA-RAM (Ours):** sellers report according to belief g_i.

4.1.2 Error Robustness Analysis

Figure 5a evaluates robustness to evaluation error (σ_{post}). The Ante-Error-Free curve is an upper bound, where sellers perfectly model h_i and best respond. **EA-RAM closely tracks this bound and maintains a stable welfare gap as σ_{post} increases**, whereas Error-Naive deteriorates sharply. Figure 5b varies seller 0's prediction error (σ_{ante}). Although welfare inevitably declines as seller information becomes noisier, **EA-RAM consistently outperforms Error-Naive**, showing that internalizing the evaluation mechanism avoids severe misallocation and yields more graceful degradation.



(a) Robustness to evaluation error σ_{post}. EA-RAM closely tracks the Ante-Error-Free upper bound, while Error-Naive degrades sharply.



(b) Robustness to prediction error σ_{ante}. EA-RAM consistently outperforms Error-Naive as seller information quality deteriorates.

Figure 5: Robustness to Dual Error. EA-RAM remains stable under evaluator and seller noise, avoiding the misallocation caused by naive reporting. Shaded regions show 95% confidence intervals.

4.2 Real-World Experiments

In this section, we benchmark EA-RAM against centralized baselines on real-world benchmarks.

4.2.1 Experimental Setup

Benchmark and baselines. We evaluate routing policies on RouterBench [17], which aggregates per-query accuracy and monetary cost for 11 LLMs (Appendix A.1), spanning both open-source (e.g., Llama, Mixtral) and proprietary models (e.g., GPT, Claude). We use queries from six benchmarks covering reasoning (HellaSwag [18], Winogrande [19], ARC-Challenge [20]), coding (MBPP [21]), math (GSM8k [22]), and knowledge-intensive understanding (MMLU [23]), and split them into train/test with a 70/30 ratio. We compare EA-RAM against a diverse set of strong baselines: EmbedLLM [4], IRT-Router [16], RouteLLM [5], FrugalGPT [24], and Cascade Routing [25]. They cover a wide range of architectures, enabling a comprehensive comparison.

Implementation details. EA-RAM trains two types of models. Each LLM has a seller-side ex-ante predictor, trained separately to estimate the probability that it will correctly answer a query based on the query embedding. The buyer maintains an ex-post evaluator that predicts answer correctness from concatenated query and answer embeddings. Both modules are MLPs with two layers. The embedding model is `all-MiniLM-L6-v2` [26]. More details are provided in Appendix A.3.

4.2.2 Routing Pareto Analysis

We evaluate EA-RAM on real-world benchmarks by sweeping each router's operating points to obtain cost–performance pairs (c, a), where c is the average cost per query and a is empirical performance, and then extracting the Pareto frontier. Beyond the base setting, EA-RAM uses two forms of seller-side local information. For oracle local information, we set $\hat{p}_m^{(\pi)}(x) = (1-\pi)\hat{p}_m(x) + \pi y_m(x)$, where $\pi \in [0,1]$ and $y_m(x)$ is the ground-truth label. For realistic local information, we retrieve the top-10 nearest training examples in the embedding space and use the similarity-weighted label average as a noisy provider-side signal $\tilde{y}_m(x)$, defining $\hat{p}_m^{(\omega)}(x) = (1-\omega)\hat{p}_m(x) + \omega\tilde{y}_m(x)$, where $\omega \in [0,1]$.

Method	MBPP	ARC-C	HellaSwag	GSM8k	MMLU	Wino.
EmbedLLM	.77453	.89855	.89646	.64596	.78289	.79154
IRT-Router	.60465	.89391	.87664	.63717	.77604	.79065
RouteLLM	.67213	.89498	.85518	.63841	.77615	.78825
FrugalGPT	.70161	.77691	.72110	.65646	.67609	.67387
Cascade Routing	.58511	.68921	.62885	.60751	.54454	.55891
EA-RAM	**.80478**	**.90353**	**.89676**	**.67302**	**.78548**	**.79396**
with π=0.1	.84932	.96729	.96919	.71379	.90021	.97214
with π=0.2	.85792	.97704	.98029	.73137	.94303	.99792
with ω=0.1	.81834	.91131	.89938	.67421	.78616	.81110
with ω=0.2	.82138	.91338	.89912	.67515	.78899	.81303

Table 2: Quantitative comparison (AIQ↑). EA-RAM yields the best overall performance.

Figure 6 and Table 2 show that EA-RAM achieves the best overall routing performance. Table 2 reports *Average Improvement in Quality* (AIQ; Appendix A.2), which averages Pareto-frontier quality over a shared cost range, with higher values indicating a better cost–performance trade-off. **EA-RAM already outperforms baselines in the base setting, and seller-side local information further shifts**

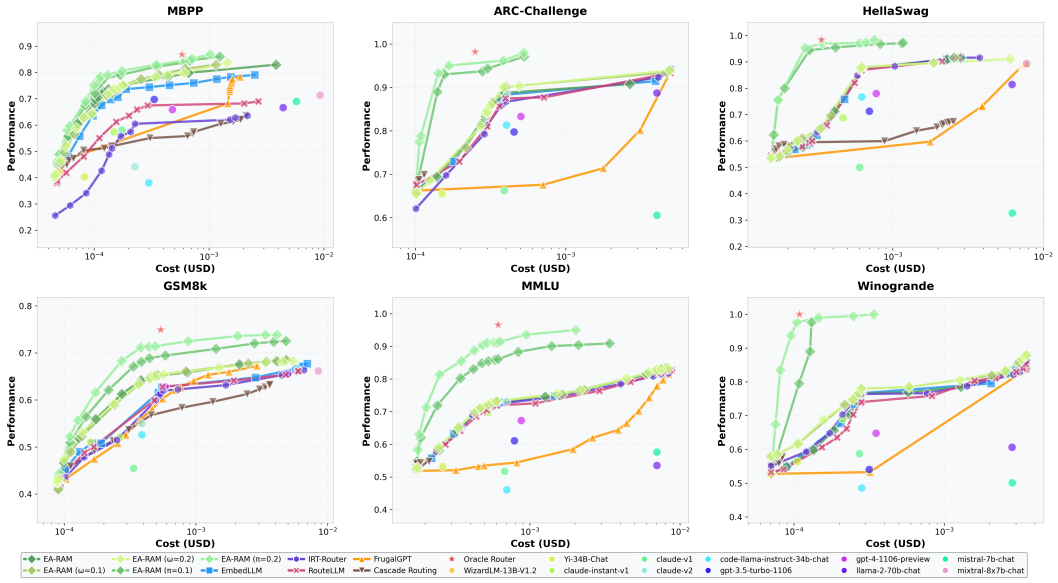
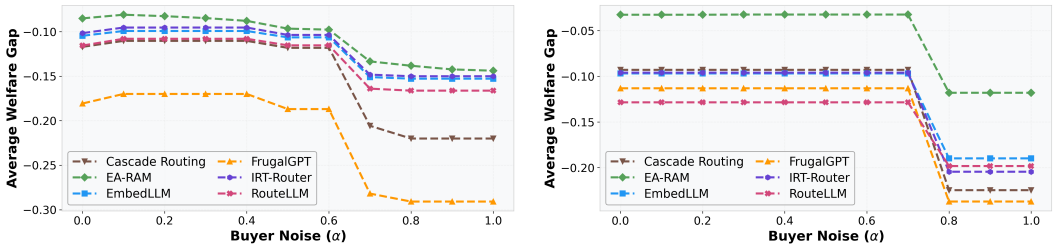


Figure 6: Cost–performance Pareto frontiers on real-world benchmarks. As π or ω increases, EA-RAM shifts the frontier upward and leftward, indicating a more favorable cost–quality trade-off than the centralized baselines.



(a) Comparison on GSM8k.

(b) Comparison on MBPP.

Figure 7: Noisy LLM-as-a-Judge evaluation. Varying the judge weight α, we report the average welfare gap to the error-free upper bound (higher is better). EA-RAM consistently incurs the smallest welfare loss.

the frontier: oracle information yields the largest gains, and realistic information also improves over the base setting across all datasets.

4.2.3 Robustness to Real-world Noise

To further test robustness to evaluator-side noise, we run a GSM8k/MBPP experiment with a LLM-as-a-judge evaluator (DeepSeek-V3.2). The evaluation score is $\text{score} = \alpha \cdot \text{score}_{\text{judge}} + (1-\alpha) \cdot \text{score}_{\text{groundtruth}}$, where $\alpha \in [0,1]$ controls evaluator noise, and score ≥ 0.5 indicates acceptance. All methods follow the same pipeline: select an LLM, evaluate its response, and compute welfare. We report the average welfare gap to the error-free upper bound over all queries, where higher is better. Figure 7 shows that **EA-RAM consistently performs best across all α, incurring the smallest welfare loss**, providing real-benchmark evidence of robustness to evaluator-side error.

4.2.4 Scalability and Efficiency

To evaluate the practical scalability and efficiency of EA-RAM, we conduct an MBPP experiment as the number of candidate LLMs increases from 3 to 7 to 11, and examine both center-side computation latency and the extra communication overhead introduced by the reverse-auction.

Center-side latency. Figure 8 shows that **the center-side latency of EA-RAM remains nearly constant as the model pool grows**, since the center only performs evaluator and auction computation and does not require retraining or per-model prediction. In contrast, the centralized router (EmbedLLM) becomes substantially slower as the model pool grows due to predictor retraining and ex-ante prediction. Detailed numerical results are provided in Appendix B.1.

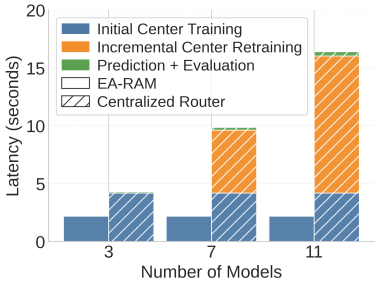


Figure 8: Center-side latency.

Communication overhead. The efficiency gain introduces extra reverse-auction communication: for each query, EA-RAM sends the query to all N providers and collects N bids, adding $(N-1)S_q + NS_b$ bytes, where S_q and S_b denote the query and bid sizes, respectively. This counts only auction-specific overhead, excluding costs shared by standard routing frameworks. As shown in Table 3, communication grows roughly linearly with N, while computation latency stays nearly constant. With parallel provider channels, communication becomes the bottleneck only below 0.338–0.384 MB/s per channel, suggesting that **it is less likely to dominate in typical high-bandwidth text-routing settings.**

N	Extra communication volume (bytes)	Computation latency (s)	Throughput threshold (MB/s)
3	44,643	0.044	0.338
7	111,439	0.043	0.370
11	177,377	0.042	0.384

Table 3: Extra reverse-auction communication overhead on MBPP.

5 Related Work

5.1 LLM Routing

To balance performance and cost, prior works have explored various routing architectures. Predictor-based approaches train a centralized router to dispatch queries, ranging from similarity-weighted ranking and embedding-based routing [5, 4] to specialized fine-tuned agents [27]. Recent works also incorporate in-context learning to adapt to new models without retraining [6]. Cascading strategies, in contrast, adopt a model chain [24, 25], invoking stronger models only when cheaper options fail. Despite these advancements, existing strategies predominantly rely on a *centralized estimation paradigm*, which places the prediction burden on the router, the party with the least internal visibility, resulting in inherent information asymmetry and scalability bottlenecks.

5.2 Market-Based Mechanisms for LLMs

Recent literature has applied auction theory to the allocation of LLM resources. Works in this domain focus primarily on determining content streams, such as auctioning slots within the RAG context window for advertisements [28], or aggregating generative preferences via token-level and summary-level auctions [29, 30]. Regarding task execution, operational frameworks such as COALESCE [15] employ reverse auctions to outsource subtasks. However, akin to classical FTMD [13], these approaches typically operate under idealized assumptions: they presume providers make perfect ex-ante predictions and the center performs perfect ex-post evaluation. Our work challenges this premise by explicitly modeling the Dual Error inherent in practical routing, and ensures incentive compatibility and optimal allocation under Dual Error.

6 Conclusion

We presented EA-RAM, an error-aware reverse-auction mechanism for LLM routing that replaces centralized capability estimation with provider-side ex-ante prediction and platform-side ex-post evaluation. This paradigm resolves the information–risk mismatch and removes the per-model profiling bottleneck by requiring only a model-agnostic evaluator. Theoretically, we model LLM routing as a Dual-Error environment with noisy provider predictions and noisy platform evaluations, and prove that EA-RAM remains incentive-aligned: it is Bayesian incentive compatible and individually rational, admits sufficient conditions for center rationality, and enjoys an explicit welfare-loss bound. Our analysis further reveals three robustness effects: opposite-signed errors can cancel, vanishing-tail links stabilize clear-cut cases via saturation, and additional noise smooths belief maps, weakly reducing the gains from marginal manipulation. Empirically, simulations and real-world benchmarks show

that EA-RAM is robust to the Dual Error and achieves a better cost–performance frontier than strong centralized baselines, with additional gains when providers contribute local information. Overall, our results suggest that market-based routing offers a scalable and robust foundation for orchestrating heterogeneous LLM ecosystems under realistic uncertainty.

References

[1] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

[2] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In *IJCAI*, 2024.

[3] Haolong Chen, Hanzhi Chen, Zijian Zhao, Kaifeng Han, Guangxu Zhu, Yichen Zhao, Ying Du, Wei Xu, and Qingjiang Shi. An overview of domain-specific foundation model: key technologies, applications and challenges. *Science China Information Sciences*, 69(1):111301, 2026.

[4] 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.

[5] 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.

[6] Chenxu Wang, Hao Li, Yiqun Zhang, Linyao Chen, Jianhao Chen, Ping Jian, Peng Ye, Qiaosheng Zhang, and Shuyue Hu. Icl-router: In-context learned model representations for llm routing. *arXiv preprint arXiv:2510.09719*, 2025.

[7] OpenRouter. Auto router. `https://openrouter.ai/docs/guides/routing/routers/auto-router`. Accessed: 2026-05-06.

[8] Requesty. Smart llm routing. `https://www.requesty.ai/solution/llm-routing`. Accessed: 2026-05-06.

[9] Shuai Yuan, Jun Wang, and Xiaoxue Zhao. Real-time bidding for online advertising: measurement and analysis. In *Proceedings of the seventh international workshop on data mining for online advertising*, pages 1–8, 2013.

[10] Jun Wang, Weinan Zhang, Shuai Yuan, et al. Display advertising with real-time bidding (rtb) and behavioural targeting. *Foundations and Trends® in Information Retrieval*, 11(4-5):297–435, 2017.

[11] Peter Cramton. The fcc spectrum auctions: An early assessment. *Journal of Economics & Management Strategy*, 6(3):431–495, 1997.

[12] Paul Robert Milgrom. *Putting auction theory to work*. Cambridge University Press, 2004.

[13] Ryan Porter, Amir Ronen, Yoav Shoham, and Moshe Tennenholtz. Fault tolerant mechanism design. *Artificial Intelligence*, 172(15):1783–1799, 2008.

[14] Hidenori Takahashi. Strategic design under uncertain evaluations: structural analysis of design-build auctions. *The RAND Journal of Economics*, 49(3):594–618, 2018.

[15] Manish Bhatt, Ronald F Del Rosario, Vineeth Sai Narajala, and Idan Habler. Coalesce: Economic and security dynamics of skill-based task outsourcing among team of autonomous llm agents. *arXiv preprint arXiv:2506.01900*, 2025.

[16] Wei Song, Zhenya Huang, Cheng Cheng, Weibo Gao, Bihan Xu, GuanHao Zhao, Fei Wang, and Runze Wu. IRT-router: Effective and interpretable multi-LLM routing via item response theory. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15629–15644, Vienna, Austria, July 2025. Association for Computational Linguistics.

[17] 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.

[18] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics.

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

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

[21] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

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

[23] 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.

[24] Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. Featured Certification.

[25] Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for LLMs. In *Forty-second International Conference on Machine Learning*, 2025.

[26] Sentence-Transformers. all-minilm-l6-v2. `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`. Accessed: 2026-05-06.

[27] Haozhen Zhang, Tao Feng, and Jiaxuan You. Router-r1: Teaching llms multi-round routing and aggregation via reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[28] MohammadTaghi Hajiaghayi, Sébastien Lahaie, Keivan Rezaei, and Suho Shin. Ad auctions for llms via retrieval augmented generation. *Advances in Neural Information Processing Systems*, 37:18445–18480, 2024.

[29] Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, pages 144–155, 2024.

[30] Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. Auctions with llm summaries. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 713–722, 2024.

Contents

A More Implementation Details

A.1 RouterBench

The model pool of the selected part of the RouterBench dataset is:

- **Open Source Models:** Llama-70B-chat, Mixtral-8x7B-chat, Yi-34B-chat, Code Llama-34B, Mistral-7B-chat, and WizardLM-13B.
- **Proprietary Models:** GPT-4, GPT-3.5-turbo, Claude-instant-v1, Claude-v1, Claude-v2.

A.2 AIQ Computation

Given a routing system family, we sample a set of parameterized routers and obtain a collection of cost–quality points $\{(c_i, q_i)\}_{i=1}^{n}$, where c_i is the average monetary cost per query and q_i is the corresponding quality (e.g., accuracy / performance). We then construct the *non-decreasing convex hull* frontier $R_f(c)$ over a shared domain $[c_{\min}, c_{\max}]$. AIQ is defined as the average quality of this frontier over the shared cost interval:

$$\mathrm{AIQ}(R_f) = \frac{1}{c_{\max} - c_{\min}} \int_{c_{\min}}^{c_{\max}} R_f(c)\, dc. \tag{1}$$

In implementation, we evaluate the integral numerically by (i) aligning all methods to the same $[c_{\min}, c_{\max}]$ via endpoint extrapolation when necessary, (ii) representing $R_f(c)$ with piecewise-linear segments, and (iii) computing the area under the curve using the trapezoidal rule, then normalizing by $(c_{\max} - c_{\min})$.

For endpoint extrapolation, let a router have observed points spanning $[c^R_{\min}, c^R_{\max}]$ with corresponding frontier quality range $[q^R_{\min}, q^R_{\max}]$. We left-extrapolate by extending the global minimum quality-cost point $(q_{\min}, c_{\min})$, and right-extrapolate by extending the router's maximum quality $q^R_{\max}$ to the global maximum cost $c_{\max}$ (i.e., $R_f(c) = q^R_{\max}$ for $c \in [c^R_{\max}, c_{\max}]$).

A.3 Other Details

EA-RAM's predictors and evaluator are trained with cross-entropy loss and AdamW for 100 epochs, with a batch size of 256 and a learning rate of 10^{-3}. All experiments were done on a server with 8 NVIDIA GeForce RTX 3090 GPUs.

B More Experiments

B.1 Detailed Results for Efficiency Analysis

We provide detailed numerical results in Table 4.

Table 4: Center-side latency under different numbers of candidate LLMs on MBPP. Lower is better.

Latency(s)/Method	EA-RAM (3)	EA-RAM (7)	EA-RAM (11)	Centralized Router (3)	Centralized Router (7)	Centralized Router (11)
Initial Center Training	**2.184**	**2.184**	**2.184**	4.181	4.181	4.181
Incremental Center Retraining	–	**0.000**	**0.000**	–	5.425	11.836
Center Computation (Prediction + Evaluation)	**0.044**	**0.043**	**0.042**	0.118	0.249	0.383
Total Latency	**2.228**	**2.227**	**2.226**	4.299	9.855	16.400

C Proof

C.1 Proof of Error-Free Setting's Property

Proposition C.1 (Optimality of the Error-Free Setting. Restatement of Proposition 3.1)**.** *In the error-free setting, the mechanism satisfies all four desirable properties: DSIC, IR, CR, and EE.*

Proof. **DSIC:** Fix others' reports. Let $H = \max\{0, \max_{k \neq i}(\hat{p}_k V - \hat{c}_k)\}$. If i wins, then $r_i = V\mu - H$, so $\mathbf{E}[r_i] = p_i V - H$ and

$$\mathbf{E}[U_i^{\text{seller}} \mid \text{win}] = \mathbf{E}[r_i] - c_i = p_i V - H - c_i = s_i - H. \tag{2}$$

If i loses, $U_i^{\text{seller}} = 0$. Thus

$$\mathbf{E}[U_i^{\text{seller}} \mid \hat{\theta}_i] = (s_i - H) \cdot \Pr[\hat{s}_i \geq H], \tag{3}$$

where the win probability depends on $\hat{s}_i$ relative to H.

- If $s_i > H$, the best outcome is to win, yielding $s_i - H > 0$. Truthful reporting sets $\hat{s}_i = s_i$, hence ensures winning. - If $s_i < H$, any win would yield $s_i - H < 0$, strictly worse than losing. Truthful reporting sets $\hat{s}_i = s_i < H$, hence ensures losing. - If $s_i = H$, both win and loss yield expected utility 0.

Thus, truth-telling maximizes expected utility in all cases.

IR: From DSIC, if i wins then $\mathbf{E}[U_i^{\text{seller}}] = s_i - H \geq 0$ because allocation requires $\hat{s}_i = s_i \geq H$. If i loses then $\mathbf{E}[U_i^{\text{seller}}] = 0$. Hence expected utility is never negative.

CR: If seller j is assigned, then

$$U^{\text{buyer}} = V\mu - r_j = V\mu - (V\mu - H) = H. \tag{4}$$

Since the dummy seller has a score $s_0 = 0$, we have $H \geq 0$, so the buyer never loses. If no seller is assigned, then $U^{\text{buyer}} = 0$.

EE: With truth-telling, $\hat{s}_i = s_i$. The mechanism assigns only if $\max_i s_i > 0$, and then to an i maximizing s_i. This maximizes expected welfare $\max\{0, \max_i s_i\}$. □

C.2 Strategic Properties

Theorem C.2 (Bayesian Incentive Compatibility (BIC). Restatement of Theorem 3.2)**.** *Seller i's interim expected utility $U_i^{seller}(\hat{s}_i)$ is maximized at $\hat{s}_i = \bar{T}_i$. Hence, reporting $\hat{s}_i = \bar{T}_i$ is a Bayesian best response. Moreover, if $\bar{T}_i > 0$, then $U_i^{seller}(\hat{s}_i)$ is uniquely maximized at $\hat{s}_i = \bar{T}_i$.*

Proof. Recall the definition of $U_i^{\text{seller}}(\hat{s}_i)$:

$$\begin{aligned} U_i^{\text{seller}}(\hat{s}_i) &= \Pr(H \leq \hat{s}_i) \cdot \mathbf{E}\big[V\tilde{\mu}_i - H - c_i \mid H \leq \hat{s}_i\big] \\ &= F_H(\hat{s}_i)\big(\bar{T}_i - \mathbf{E}[H \mid H \leq \hat{s}_i]\big). \end{aligned} \tag{5}$$

Differentiating $U_i^{\text{seller}}(\hat{s}_i)$ and using

$$\frac{d}{dx}\big[F_H(x)\mathbf{E}(H \mid H \leq x)\big] = f_H(x)\, x \tag{6}$$

gives

$$U_i^{\text{seller}'}(\hat{s}_i) = f_H(\hat{s}_i)\,(\bar{T}_i - \hat{s}_i). \tag{7}$$

If $\bar{T}_i > 0$, then $U_i^{\text{seller}'}(\hat{s}_i) > 0$ for $\hat{s}_i < \bar{T}_i$ and $U_i^{\text{seller}'}(\hat{s}_i) < 0$ for $\hat{s}_i > \bar{T}_i$, so $U_i^{\text{seller}}(\hat{s}_i)$ is uniquely maximized at $\hat{s}_i = \bar{T}_i$.

If $\bar{T}_i \leq 0$, then for every $\hat{s}_i > 0$ we have $U_i^{\text{seller}'}(\hat{s}_i) < 0$, so $U_i^{\text{seller}}(\hat{s}_i)$ is decreasing on $(0, \infty)$. Hence, any positive over-report is not profitable. With the explicit null option, any non-positive report is equivalent to opting out, so reporting $\hat{s}_i = \bar{T}_i$ remains optimal. Therefore, $\hat{s}_i = \bar{T}_i$ is always a Bayesian best response, and uniqueness holds whenever $\bar{T}_i > 0$. □

Theorem C.3 (Individual Rationality (IR). Restatement of Theorem 3.3)**.** *At the equilibrium $\hat{s} = \bar{T}$, every seller satisfies IR: $\mathbf{E}[U_i^{seller} \mid \tilde{\theta}_i] \geq 0$.*

Proof. **Win case.** Conditioned on $H \leq \bar{T}_i$,

$$\mathbf{E}[U_i^{\text{seller}} \mid \text{win}, \tilde{\theta}_i] \;=\; \bar{T}_i - \mathbf{E}[H \mid H \leq \bar{T}_i] \;\geq\; 0. \tag{8}$$

Plugging $\hat{s}_i = \bar{T}_i$ into the interim utility $U_i^{\text{seller}}(\hat{s}_i)$ yields

$$U_i^{\text{seller}}(\bar{T}_i) = F_H(\bar{T}_i)\big(\bar{T}_i - \mathbf{E}[H \mid H \leq \bar{T}_i]\big) \;\geq\; 0. \tag{9}$$

Lose case. If $H > \bar{T}_i$, then

$$\mathbf{E}[U_i^{\text{seller}} \mid \text{lose}, \tilde{\theta}_i] = 0. \tag{10}$$

Combining (8)–(10) gives $\mathbf{E}[U_i^{\text{seller}} \mid \tilde{\theta}_i] \geq 0$. □

Lemma C.4 (Lipschitz Bound.)**.** *The link function σ is globally L_σ-Lipschitz, i.e., $|\sigma'(z)| \le L_\sigma$ for all $z \in \mathbb{R}$. Therefore, for any real number x and any random variable ϵ with $\mathbf{E}[\epsilon^2] < \infty$, the deviation between $\mathbf{E}[\sigma(x+\epsilon)]$ and $\sigma(x)$ satisfies $\big|\mathbf{E}[\sigma(x+\epsilon)] - \sigma(x)\big| \le L_\sigma\, \mathbf{E}[|\epsilon|] \le L_\sigma \sqrt{\mathrm{Var}(\epsilon) + (\mathbf{E}[\epsilon])^2}$.*

Proof. The proof consists of two steps: applying the Lipschitz property to the expectation and then bounding the first absolute moment using the second moment.

Step 1: Bounding the deviation via the Lipschitz constant. Since σ is differentiable and satisfies $|\sigma'(z)| \le L_\sigma$ for all z, by the Mean Value Theorem, σ is L_σ-Lipschitz continuous. That is, for any $a, b \in \mathbb{R}$, $|\sigma(a) - \sigma(b)| \le L_\sigma |a - b|$. Let $a = x + \epsilon$ and $b = x$. Then:

$$|\sigma(x+\epsilon) - \sigma(x)| \le L_\sigma |(x+\epsilon) - x| = L_\sigma |\epsilon|. \tag{11}$$

Now, consider the absolute difference of the expectations. By the linearity of expectation and Jensen's inequality (since the absolute value function $|\cdot|$ is convex, $|\mathbf{E}[Y]| \le \mathbf{E}[|Y|]$), we have:

$$\big|\mathbf{E}[\sigma(x+\epsilon)] - \sigma(x)\big| = \big|\mathbf{E}[\sigma(x+\epsilon) - \sigma(x)]\big| \tag{12}$$

$$\le \mathbf{E}\big[|\sigma(x+\epsilon) - \sigma(x)|\big]. \tag{13}$$

Substituting the Lipschitz bound from Eq. (1) into the expectation:

$$\mathbf{E}\big[|\sigma(x+\epsilon) - \sigma(x)|\big] \le \mathbf{E}[L_\sigma |\epsilon|] = L_\sigma \mathbf{E}[|\epsilon|]. \tag{14}$$

This establishes the first inequality of the lemma.

Step 2: Bounding the first moment via Variance. We apply Lyapunov's inequality (or simply Jensen's inequality for the convex function $f(y) = y^2$), which states that $(\mathbf{E}[|Y|])^2 \le \mathbf{E}[Y^2]$. Applied to the random variable ϵ:

$$\mathbf{E}[|\epsilon|] \le \sqrt{\mathbf{E}[\epsilon^2]}. \tag{15}$$

Recall the definition of variance: $\mathrm{Var}(\epsilon) = \mathbf{E}[\epsilon^2] - (\mathbf{E}[\epsilon])^2$. Rearranging for the second moment gives $\mathbf{E}[\epsilon^2] = \mathrm{Var}(\epsilon) + (\mathbf{E}[\epsilon])^2$. Substituting this into Eq. (4):

$$\mathbf{E}[|\epsilon|] \le \sqrt{\mathrm{Var}(\epsilon) + (\mathbf{E}[\epsilon])^2}. \tag{16}$$

Multiplying both sides by L_σ yields the final bound:

$$L_\sigma \mathbf{E}[|\epsilon|] \le L_\sigma \sqrt{\mathrm{Var}(\epsilon) + (\mathbf{E}[\epsilon])^2}. \tag{17}$$

□

Theorem C.5 (Center Rationality (CR). Restatement of Theorem 3.4)**.** *Each of the following sufficient conditions guarantees CR: (A) $\mathbf{E}[H] \ge V\Delta_{gate}$, where $\Delta_{gate} = L_\sigma \sqrt{b_{post} + a_{post}^2}$; (B) $\Delta_{cons} = \mathbf{E}\big[p_{(1)} - h_{(1)}\big] \ge 0$.*

Proof. Let the runner-up surplus be H, and let the selected winner be indexed by (1). The winner's true fulfillment probability and evaluated acceptance probability are

$$p_{(1)} = \sigma(\phi_{(1)}), \qquad h_{(1)} = \sigma\big(\phi_{(1)} + \varepsilon_{\text{post}}\big). \tag{18}$$

Let $\mu_{(1)} \sim \mathrm{Bernoulli}(p_{(1)})$ denote the true fulfillment outcome and $\tilde{\mu}_{(1)} \sim \mathrm{Bernoulli}(h_{(1)})$ denote the evaluator's acceptance signal. Since the winner is paid

$$r_{(1)} = V\tilde{\mu}_{(1)} - H, \tag{19}$$

the buyer's realized utility is

$$U^{\text{buyer}} = V\mu_{(1)} - r_{(1)} = H + V\big(\mu_{(1)} - \tilde{\mu}_{(1)}\big). \tag{20}$$

Taking expectations and using $\mathbf{E}[\mu_{(1)}] = \mathbf{E}[p_{(1)}]$ and $\mathbf{E}[\tilde{\mu}_{(1)}] = \mathbf{E}[h_{(1)}]$, we obtain

$$\mathbf{E}[U^{\text{buyer}}] = \mathbf{E}[H] + V\big(\mathbf{E}[p_{(1)}] - \mathbf{E}[h_{(1)}]\big). \tag{21}$$

(A) Condition on $\phi_{(1)}$ and apply Lemma C.4 to $x = \phi_{(1)}$ and $\epsilon = \varepsilon_{\text{post}}$:

$$\big|\mathbf{E}[h_{(1)} \mid \phi_{(1)}] - p_{(1)}\big| \;\le\; L_\sigma \sqrt{b_{\text{post}} + a_{\text{post}}^2}\,. \tag{22}$$

Taking expectations over $\phi_{(1)}$ gives

$$\left|\mathbf{E}[h_{(1)}] - \mathbf{E}[p_{(1)}]\right| \leq \Delta_{\text{gate}}. \tag{23}$$

Substituting (23) into (21), we obtain

$$\mathbf{E}[U^{\text{buyer}}] \geq \mathbf{E}[H] - V\Delta_{\text{gate}}. \tag{24}$$

Hence, if $\mathbf{E}[H] \geq V\Delta_{\text{gate}}$, then $\mathbf{E}[U^{\text{buyer}}] \geq 0$, establishing CR.

(B) By (21),

$$\mathbf{E}[U^{\text{buyer}}] = \mathbf{E}[H] + V\,\mathbf{E}[p_{(1)} - h_{(1)}] = \mathbf{E}[H] + V\Delta_{\text{cons}}. \tag{25}$$

Since the mechanism includes the dummy seller with score 0, we have $H \geq 0$ and thus $\mathbf{E}[H] \geq 0$. Therefore, whenever $\Delta_{\text{cons}} \geq 0$, (25) implies $\mathbf{E}[U^{\text{buyer}}] \geq 0$. This establishes CR. □

Proposition C.6 (Comparative Statics: Ability and Difficulty. Restatement of Proposition 3.5)**.** *Holding fixed the runner-up score distribution F_H, seller i's interim expected utility $\mathbf{E}[U_i^{seller}]$ is weakly increasing in their model ability m_i and weakly decreasing in the task difficulty d.*

Proof. We consider ceteris-paribus comparative statics, holding the runner-up score distribution F_H fixed. Recall that

$$\mathbf{E}[U_i^{\text{seller}}] = \int_{-\infty}^{\bar{T}_i} (\bar{T}_i - h)\, dF_H(h), \tag{26}$$

where $\bar{T}_i = Vg_i - c_i$. Differentiating with respect to x using the Leibniz integral rule yields:

$$\frac{\partial \mathbf{E}[U_i^{\text{seller}}]}{\partial x} = \Pr(H \leq \bar{T}_i) \cdot V \cdot \frac{\partial g_i}{\partial x} = \Pr(H \leq \bar{T}_i) \cdot V \cdot \mathbf{E}[\sigma'(\phi + \eta_i)]\, \frac{\partial \phi}{\partial x}. \tag{27}$$

The second equality follows from the chain rule applied to the ex-ante belief g_i. Observe that V and $\mathbf{E}[\sigma']$ are strictly positive, while the win probability $\Pr(H \leq \bar{T}_i)$ is non-negative. Thus, the sign of the utility gradient follows the sign of $\frac{\partial \phi}{\partial x}$. Recall from Section 2.1 that ϕ is non-decreasing in ability ($\frac{\partial \phi}{\partial m_i} \geq 0$) and non-increasing in difficulty ($\frac{\partial \phi}{\partial d} \leq 0$). Therefore:

1. For ability ($x = m_i$), the gradient is non-negative ($\frac{\partial \mathbf{E}[U_i^{\text{seller}}]}{\partial m_i} \geq 0$).
2. For difficulty ($x = d$), the gradient is non-positive ($\frac{\partial \mathbf{E}[U_i^{\text{seller}}]}{\partial d} \leq 0$).

This confirms that the seller's utility is weakly increasing in ability and weakly decreasing in difficulty. □

C.3 Welfare Analysis

Proposition C.7 (Economic Efficiency (EE). Restatement of Proposition 3.6)**.** *The equilibrium allocation induced by the error-aware setting attains the same expected welfare as the error-free benchmark if and only if the error-aware winner $i^\dagger$ is welfare-optimal.*

Proof. Let $i^\star$ denote an error-free welfare-maximizing seller. If seller i is selected, then the realized welfare is

$$W_i = V\mu_i - c_i. \tag{28}$$

Taking the expectation gives

$$\mathbf{E}[W_i] = V\mathbf{E}[\mu_i] - c_i = Vp_i - c_i. \tag{29}$$

Hence, the expected welfare induced by the error-aware allocation is

$$\mathbf{E}[W_{i^\dagger}] = Vp_{i^\dagger} - c_{i^\dagger}, \tag{30}$$

whereas the error-free benchmark welfare is

$$\mathbf{E}[W_{i^\star}] = \max_i \{Vp_i - c_i\}. \tag{31}$$

Therefore,

$$\mathbf{E}[W_{i^\dagger}] = \mathbf{E}[W_{i^\star}] \iff Vp_{i^\dagger} - c_{i^\dagger} = \max_i \{Vp_i - c_i\}. \tag{32}$$

This proves the claim. □

Theorem C.8 (Welfare-loss bound. Restatement of Theorem 3.7)**.** *The expected welfare loss satisfies* $0 \le \mathbf{E}[W_{i^\star}] - \mathbf{E}[W_{i^\dagger}] = \mathbf{E}[(Vp_{i^\star} - c_{i^\star}) - (Vp_{i^\dagger} - c_{i^\dagger})] \le 2VL_\sigma(M_{\text{post}} + M_{\text{ante}})$.

Proof. **Step 1: Pointwise comparison.** Let $\mathbf{E}[W_i] = Vp_i - c_i$ and $\bar{T}_i = Vg_i - c_i$. Since $i^\dagger$ maximizes $\bar{T}_i$,

$$\mathbf{E}[W_{i^\star}] - \mathbf{E}[W_{i^\dagger}] = (\mathbf{E}[W_{i^\star}] - \bar{T}_{i^\star}) + (\bar{T}_{i^\star} - \bar{T}_{i^\dagger}) + (\bar{T}_{i^\dagger} - \mathbf{E}[W_{i^\dagger}]) \;\le\; (\mathbf{E}[W_{i^\star}] - \bar{T}_{i^\star}) + (\bar{T}_{i^\dagger} - \mathbf{E}[W_{i^\dagger}]), \tag{33}$$

because $\bar{T}_{i^\star} - \bar{T}_{i^\dagger} \le 0$. Rearranging yields

$$0 \le \mathbf{E}[W_{i^\star}] - \mathbf{E}[W_{i^\dagger}] \le V\big[(p_{i^\star} - g_{i^\star}) + (g_{i^\dagger} - p_{i^\dagger})\big] \le V\big(|g_{i^\star} - p_{i^\star}| + |g_{i^\dagger} - p_{i^\dagger}|\big). \tag{34}$$

Step 2: Bound $|g_i - p_i|$ by decomposing the two channels. Introduce the telescoping decomposition

$$g_i^{(0)} = \sigma(\phi_i) = p_i, \quad g_i^{(1)} = \mathbf{E}[\sigma(\phi_i + \varepsilon_{\text{post}})], \quad g_i^{(2)} = g_i, \tag{35}$$

so that

$$|g_i - p_i| \le |g_i^{(1)} - g_i^{(0)}| + |g_i^{(2)} - g_i^{(1)}|. \tag{36}$$

Applying Lemma C.4 to each increment:

$$|g_i^{(1)} - g_i^{(0)}| \le L_\sigma \sqrt{b_{\text{post}} + a_{\text{post}}^2} = L_\sigma M_{\text{post}}, \tag{37}$$

$$|g_i^{(2)} - g_i^{(1)}| \le L_\sigma \sqrt{b_{\text{ante},i} + a_{\text{ante},i}^2} \le L_\sigma M_{\text{ante}}. \tag{38}$$

Therefore,

$$|g_i - p_i| \le L_\sigma\big(M_{\text{post}} + M_{\text{ante}}\big) \quad \text{for all } i. \tag{39}$$

Step 3: Combine. Taking expectations of (34) and applying (39) at indices $i^\star$ and $i^\dagger$,

$$0 \;\le\; \mathbf{E}[W_{i^\star}] - \mathbf{E}[W_{i^\dagger}] \;\le\; V\,\mathbf{E}\big[|g_{i^\star} - p_{i^\star}| + |g_{i^\dagger} - p_{i^\dagger}|\big] \;\le\; 2VL_\sigma\big(M_{\text{post}} + M_{\text{ante}}\big), \tag{40}$$

completing the proof. □

C.4 Structural Insights

Proposition C.9 (Opposite-Signed Error Compensation. Restatement of Proposition 3.8)**.** *If* $(g_i - h_i)(h_i - p_i) \le 0$ *for* $i \in \{i^\star, i^\dagger\}$*, then the welfare-loss bound tightens to* $2VL_\sigma \max\{M_{\text{ante}}, M_{\text{post}}\}$.

Proof. Let $i^\star \in \arg\max_j W_j$ denote the winner under the error-free setting, and let $i^\dagger \in \arg\max_j\{Vg_j - c_j\}$ denote the winner selected by the error-involved mechanism. We start from the welfare-gap decomposition established in Theorem 3.7:

$$0 \le \mathbf{E}[W_{i^\star}] - \mathbf{E}[W_{i^\dagger}] \le V\Big(|g_{i^\star} - p_{i^\star}| + |g_{i^\dagger} - p_{i^\dagger}|\Big). \tag{41}$$

Consider the error decomposition $g_i - p_i = (g_i - h_i) + (h_i - p_i) \coloneqq a_i + b_i$. The condition $(g_i - h_i)(h_i - p_i) \le 0$ implies that the two error components a_i and b_i have opposite signs. Consequently, their sum is bounded by the maximum of their absolute values:

$$|g_i - p_i| = |a_i + b_i| \le \max\{|a_i|, |b_i|\}. \tag{42}$$

Applying this to the critical sellers $i \in \{i^\star, i^\dagger\}$ yields:

$$|g_{i^\star} - p_{i^\star}| + |g_{i^\dagger} - p_{i^\dagger}| \;\le\; \max\{|a_{i^\star}|, |b_{i^\star}|\} + \max\{|a_{i^\dagger}|, |b_{i^\dagger}|\}. \tag{43}$$

Next, invoke the Lipschitz bounds from Lemma C.4:

$$|g_i - h_i| \le L_\sigma\, M_{\text{ante}}, \tag{44}$$

$$|h_i - p_i| \le L_\sigma\, M_{\text{post}}, \tag{45}$$

where the first arises because $g - h$ differs only by the ante channel, and the second because $h - p$ aggregates the post channels. Taking the supremum across sellers j yields

$$\sup_j |g_j - h_j| \le L_\sigma M_{\text{ante}}, \quad \sup_j |h_j - p_j| \le L_\sigma M_{\text{post}}. \tag{46}$$

Combining (41)–(46), we obtain

$$0 \;\le\; \mathbf{E}[W_{i^\star}] - \mathbf{E}[W_{i^\dagger}] \;\le\; 2VL_\sigma \,\max\bigl\{M_{\text{ante}},\, M_{\text{post}}\bigr\}. \tag{47}$$

Finally, because for any nonnegative x, y, $\max\{x, y\} < x + y$ when both are positive, we have

$$2VL_\sigma \,\max\bigl\{M_{\text{ante}},\, M_{\text{post}}\bigr\} \;<\; 2VL_\sigma\,(M_{\text{post}} + M_{\text{ante}}), \tag{48}$$

which completes the proof. □

Proposition C.10 (Saturation Robustness. Restatement of Proposition 3.9)**.** $\lim_{|x|\to\infty} \sigma'(x) = 0$. *For any error ϵ with $\mathbf{E}[\epsilon^2] < \infty$, the deviation $\Delta(\phi_i) = |\mathbf{E}[\sigma(\phi_i + \epsilon)] - \sigma(\phi_i)|$ satisfies* $\lim_{|\phi_i|\to\infty} \Delta(\phi_i) = 0$.

Proof. Using the integral form of the Mean Value Theorem, for any realization of ϵ,

$$\sigma(\phi_i + \epsilon) - \sigma(\phi_i) = \epsilon \int_0^1 \sigma'(\phi_i + t\epsilon)\, dt. \tag{49}$$

Taking absolute values and expectations gives

$$\Delta(\phi_i) \le \mathbf{E}\left[|\epsilon| \int_0^1 |\sigma'(\phi_i + t\epsilon)|\, dt\right]. \tag{50}$$

Fix $\eta > 0$. Since $\lim_{|x|\to\infty} \sigma'(x) = 0$, there exists $R > 0$ such that

$$|x| \ge R \implies |\sigma'(x)| \le \eta. \tag{51}$$

Define the event $A = \{|\epsilon| \le |\phi_i|/2\}$ and split the expectation in (50) over A and A^c.

On A, for any $t \in [0, 1]$,

$$|\phi_i + t\epsilon| \ge |\phi_i| - t|\epsilon| \ge |\phi_i| - \frac{|\phi_i|}{2} = \frac{|\phi_i|}{2}. \tag{52}$$

Hence, if $|\phi_i| \ge 2R$, then $|\phi_i + t\epsilon| \ge R$ and by (51) we have $|\sigma'(\phi_i + t\epsilon)| \le \eta$. Therefore,

$$\mathbf{E}\left[|\epsilon| \int_0^1 |\sigma'(\phi_i + t\epsilon)|\, dt\; \mathbf{1}_A\right] \le \eta\, \mathbf{E}[|\epsilon|]. \tag{53}$$

On A^c, we use boundedness of σ':

$$\mathbf{E}\left[|\epsilon| \int_0^1 |\sigma'(\phi_i + t\epsilon)|\, dt\; \mathbf{1}_{A^c}\right] \le \|\sigma'\|_\infty\, \mathbf{E}[|\epsilon|\, \mathbf{1}\{|\epsilon| > |\phi_i|/2\}]\,. \tag{54}$$

Since $\mathbf{E}[\epsilon^2] < \infty$, we have $\mathbf{E}[|\epsilon|] < \infty$, and moreover

$$\mathbf{E}[|\epsilon|\, \mathbf{1}\{|\epsilon| > t\}] \xrightarrow[t\to\infty]{} 0 \tag{55}$$

(e.g., by dominated convergence with the dominating integrable random variable $|\epsilon|$). Thus the right-hand side of (54) vanishes as $|\phi_i| \to \infty$.

Combining (50)–(54) yields, for all sufficiently large $|\phi_i|$,

$$\Delta(\phi_i) \le \eta\, \mathbf{E}[|\epsilon|] + \|\sigma'\|_\infty\, \mathbf{E}[|\epsilon|\, \mathbf{1}\{|\epsilon| > |\phi_i|/2\}]\,. \tag{56}$$

Taking $\limsup_{|\phi_i|\to\infty}$ gives $\limsup_{|\phi_i|\to\infty} \Delta(\phi_i) \le \eta\, \mathbf{E}[|\epsilon|]$. Because $\eta > 0$ is arbitrary, we conclude that $\Delta(\phi_i) \to 0$ as $|\phi_i| \to \infty$. □

Proposition C.11 (Noise-Induced Flattening. Restatement of Proposition 3.10)**.** *Let η be an independent noise term with $\mathbf{E}[|\eta|] < \infty$. Define the perturbed belief maps by convolution: $\tilde{g}_i(\phi) = \mathbf{E}[g_i(\phi + \eta)]$ and $\tilde{h}_i(\phi) = \mathbf{E}[h_i(\phi + \eta)]$. If g_i and h_i are continuously differentiable with bounded derivatives, then* $\sup_\phi |\frac{\partial \tilde{g}_i}{\partial \phi}(\phi)| \le \sup_\phi |\frac{\partial g_i}{\partial \phi}(\phi)|$ *and* $\sup_\phi |\frac{\partial \tilde{h}_i}{\partial \phi}(\phi)| \le \sup_\phi |\frac{\partial h_i}{\partial \phi}(\phi)|$. *Consequently, injecting additional independent noise weakly reduces the maximal sensitivity of both $\phi \mapsto g_i(\phi)$ and $\phi \mapsto h_i(\phi)$.*

Proof. We prove the claim for $\tilde{g}_i$; the argument for $\tilde{h}_i$ is identical. By definition,

$$\tilde{g}_i(\phi) = \mathbf{E}[g_i(\phi + \eta)]. \tag{57}$$

Since g_i is continuously differentiable and g_i' is bounded, we may differentiate under the expectation (e.g., by dominated convergence) to obtain

$$\frac{\partial \tilde{g}_i}{\partial \phi}(\phi) = \mathbf{E}[g_i'(\phi + \eta)]\,. \tag{58}$$

Taking absolute values and using the bound $|g_i'(x)| \le \sup_y |g_i'(y)|$ for all x yields

$$|\frac{\partial \tilde{g}_i}{\partial \phi}(\phi)| = |\mathbf{E}[g_i'(\phi + \eta)]| \le \mathbf{E}[|g_i'(\phi + \eta)|] \le \sup_y |g_i'(y)|. \tag{59}$$

Finally, taking the supremum over ϕ on the left-hand side gives

$$\sup_\phi |\tfrac{\partial \tilde{g}_i}{\partial \phi}(\phi)| \le \sup_y |g_i'(y)|, \tag{60}$$

as desired. The same steps apply to $\tilde{h}_i$. □

D Limitations and Future Directions

Although this work, to our knowledge, is the first to introduce a reverse-auction paradigm for LLM routing and to explicitly model Dual Error in this setting, several limitations remain. Our present theory does not cover payment-rule variants such as bounded penalties or non-negative payments, which may alter the existing BIC, IR, and CR guarantees. Although we provide robustness experiments with realistic noisy local information and noisy evaluators, our empirical evaluation still does not cover all practical settings. Moreover, while EA-RAM improves scalability by trading additional communication overhead for nearly fixed center-side computation, this advantage is most evident in our current text-based benchmarks, where payload sizes are relatively small; in communication-heavy settings such as image- or video-based tasks, or under degraded network conditions, communication may become a more significant bottleneck. In addition, the auction protocol requires broadcasting each query to all candidate providers before selection, which may expose user prompts to non-winning providers and raises security concerns. Finally, because routing and payment are tied to evaluator outcomes, providers may be incentivized to optimize toward the evaluator rather than the user's true underlying need.

These limitations suggest several directions for future work. An important extension is to develop communication-efficient reverse-auction mechanisms. Other promising directions include extending the theory to bounded-penalty or non-negative-payment variants, tightening the current welfare-loss bound, generalizing the framework beyond single-winner routing to support multi-provider cross-checking, cascading, and multi-turn settings, extending the model to heterogeneous risk preferences, and providing more principled support for non-binary evaluation tasks beyond the current threshold-based implementation. In addition, while EA-RAM is defined with respect to an announced task value V, it remains well defined as long as the buyer announces a reference value; a natural next step is to study more systematically how misspecification of V changes participation incentives and the resulting operating point.