ERROR-AWARE REVERSE AUCTION MECHANISM FOR LARGE LANGUAGE MODEL ROUTING

Anonymous authors
Paper under double-blind review

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 formulate LLM routing as a market-based allocation problem among strategic providers and propose a routing paradigm that shifts ex-ante prediction to LLM providers via a reverse auction, where providers submit self-predicted acceptance probabilities and execution costs. To account for 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 Dual Error. We prove that, under a private-evaluation-belief structure, truthful effective-surplus reporting is incentive compatible in the reduced-form score space and individually rational under sellers' subjective beliefs, establish sufficient conditions for center rationality, and derive an explicit social-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 and reduces their maximal local sensitivity. Simulations and real-world benchmarks show that EA-RAM is robust to Dual Error and achieves a better cost–performance Pareto frontier than centralized baselines, with additional gains from provider-side local information, validating its practical effectiveness.

1 INTRODUCTION

Large language models (LLMs) underpin many intelligent applications (Guo et al., 2024; Chen et al., 2026a;b; Li et al., 2026), 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.

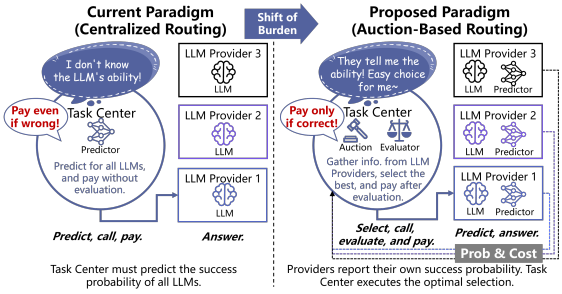


Figure 1: Paradigm shift from traditional Centralized Routing to our proposed Auction-Based Routing.

Most existing routers follow a centralized estimation paradigm, where the task center predicts each model's performance (Zhuang et al., 2025; Zhang et al., 2026a). 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 (Wang et al., 2025), which hinders rapid expansion.

To address these, 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 submit surplus scores formed from their local performance estimates and execution costs; the center maintains only a model-agnostic ex-post evaluator. This distributed

[1] e.g., OpenRouter and Requesty

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Conceptual comparison of standard VCG, FTMD, and EA-RAM. EA-RAM retains an execution-contingent externality-payment structure but differs in the information governing allocation, settlement, and welfare.

	Standard VCG (Groves, 1973)	**FTMD** (Porter et al., 2008)	**EA-RAM**
Modeled uncertainty	Standard private values	Execution uncertainty	Dual Error
Decision-relevant value	True private value	$Vp_i - c_i$	$Vg_i - c_i$
Settlement basis	Reported-value externality	True execution outcome μ_i	Evaluator signal $\tilde{\mu}_i$
Information alignment	Valuation governs incentives and welfare	p_i governs incentives and welfare	g_i, h_i, and p_i govern incentives, settlement, and welfare
Incentive guarantee	DSIC in value space	DSIC in execution-surplus space	DSIC in reduced-form effective-surplus space
Efficiency guarantee	Exact	Exact under observable execution outcomes	Bounded welfare loss under Dual Error

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 (Yuan et al., 2013; Wang et al., 2017) and auction-based spectrum allocation (Cramton, 1997; Milgrom, 2004). However, transferring this idea to LLM routing is *not* straightforward. Analyzing such a mechanism requires extending classical execution-contingent mechanism design to an inherent Dual-Error information structure: providers' ex-ante beliefs are noisy, while the center's ex-post evaluation is imperfect. This setting departs from prior fault-tolerant allocation mechanisms (Porter et al., 2008; Takahashi, 2018; Bhatt et al., 2025), 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, under Dual Error and a private-evaluation-belief structure, EA-RAM admits dominant-strategy truthful effective-surplus reporting in the reduced-form score space and is individually rational (IR) under sellers' subjective beliefs; 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 noise smooths belief maps and reduces their maximal local sensitivity. 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 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. EA-RAM is DSIC in the reduced-form effective-surplus space under private evaluation beliefs, satisfies IR, admits sufficient conditions for CR, and enjoys a welfare-loss bound; it further reveals 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 cost–performance trade-offs compared to centralized baselines.

2 Related Work

Market-Based Mechanisms for LLMs. Auction-based methods have been studied for LLM content allocation (Hajiaghayi et al., 2024; Duetting et al., 2024; Dubey et al., 2024) and task outsourcing (Bhatt et al., 2025; Habibi & Ercetin, 2025; Xu et al., 2026; Fan et al., 2026). While mechanisms such as FTMD (Porter et al., 2008) handle execution uncertainty via execution-contingent transfers, EA-RAM studies a *Dual-Error* regime where fulfillment, center evaluation, and seller belief may differ, changing the strategic type, center rationality, and welfare guarantees (Table 1).

LLM Routing. Existing LLM routers mainly rely on centralized performance estimation, including predictor-based routing (Song et al., 2025; Ong et al., 2025; Zhuang et al., 2025; Zhang et al., 2026c;a; Niu et al., 2026; Zhang et al., 2026b), in-context adaptation (Wang et al., 2025), and cascad-

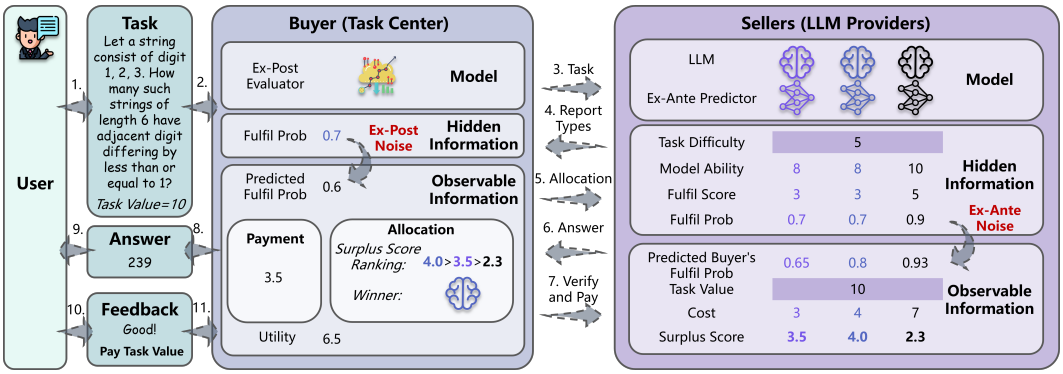


Figure 2: EA-RAM for LLM routing. (1) Providers bid using ex-ante predictions. (2) The buyer allocates the query by reported surplus. (3) The selected model answers. (4) The buyer evaluates the output and settles payment. (5) The model output is returned to the user.

ing (Chen et al., 2024; Dekoninck et al., 2025). However, 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.

3 Error-Aware Reverse Auction Mechanism

3.1 Basic Setting

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. 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 task has a common-knowledge *value* $V > 0$, realized by the buyer iff the task demand is fulfilled ($\mu = 1$). Seller i is characterized by the physical primitives (p_i, c_i), where $c_i > 0$ is its execution cost and p_i is the true fulfillment probability of the binary fulfillment indicator $\mu_i \sim \text{Bernoulli}(p_i)$. These quantities characterize the underlying welfare environment and are distinct from the seller's strategic type introduced below.

Inspired by capability modeling (Wang et al., 2025; Song et al., 2025), we model *fulfillment probability* p_i via the alignment between seller i's *model ability* $m_i \in \mathbb{R}$ and *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 *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.

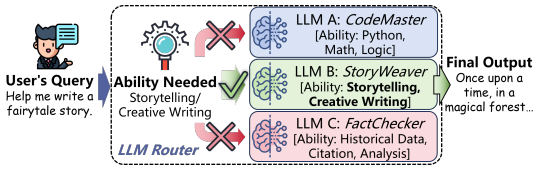


Figure 3: Ability–Difficulty Matching.

3.2 Error-Aware Prediction & Evaluation

LLM Routing faces *Dual Error* due to noisy prediction and imperfect evaluation, undermining classical mechanisms such as FTMD (Porter et al., 2008) and necessitating an error-aware formulation.

Buyer's ex-post evaluation. The buyer's evaluator may deviate from the ground-truth fulfillment criterion. We model this evaluator-side uncertainty by ϵ_{post}, with $\mathbb{E}[\epsilon_{\text{post}}] = a_{\text{post}}$ and $\text{Var}(\epsilon_{\text{post}}) = b_{\text{post}}$.

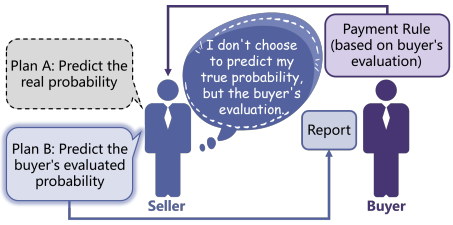


Figure 4: Seller's strategic decision. The seller bids based on the buyer's error-involved evaluation, rather than the unobserved ground-truth.

[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.

The buyer's marginal acceptance probability is $h_i = \mathbb{E}_{\epsilon_{\text{post}}}[\sigma(\phi_i + \epsilon_{\text{post}})]$. The resulting evaluation signal satisfies $\tilde{\mu}_i \sim \text{Bernoulli}(h_i)$. The evaluator and settlement rule are publicly committed before bidding and remain fixed throughout the auction. Thus, the buyer cannot alter the evaluation rule after observing bids or model outputs.

Sellers' ex-ante prediction. Since payment depends on the buyer's evaluation, seller i forms a subjective belief about the buyer's acceptance criterion rather than the unobserved ground-truth fulfillment probability. Let $\epsilon_{\text{ante},i}$ denote seller i's belief error about the buyer's evaluation criterion, independent of ϵ_{post}, and define $\eta_i = \epsilon_{\text{post}} + \epsilon_{\text{ante},i}$. The seller's subjective acceptance probability is $g_i = \mathbb{E}_{\eta_i}[\sigma(\phi_i + \eta_i)]$, where $\mathbb{E}[\eta_i] = a_{\text{post}} + a_{\text{ante},i}$ and $\text{Var}(\eta_i) = b_{\text{post}} + b_{\text{ante},i}$. Thus, ϵ_{post} captures the buyer's evaluation distortion relative to ground truth, whereas $\epsilon_{\text{ante},i}$ captures seller i's uncertainty about that evaluation criterion.

3.3 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. The latent variables p_i and μ_i are used only for welfare analysis and need not be observed by the mechanism.

Algorithm 1 EA-RAM

Input: Sellers $\mathcal{I} = \{1, \dots, N\}$; task (d, V).
for $i = 1, \dots, N$ **do**
  Seller i privately forms its evaluation belief g_i and submits $\hat{\theta}_i = (\hat{p}_i, \hat{c}_i)$.
  Set the induced score $\hat{s}_i \leftarrow V\hat{p}_i - \hat{c}_i$.
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
Seller j executes the task and returns its output.
The buyer applies the committed evaluator and observes $\tilde{\mu}_j$.
Pay $r_j \leftarrow V\tilde{\mu}_j - H$; set $r_i \leftarrow 0$ for all $i \neq j$.
return $j, \{r_i\}_{i=1}^N$.

Utilities and welfare. The buyer aims to maximize expected social welfare, while each seller maximizes its own expected utility under the mechanism. A non-winning seller receives zero utility, whereas the winner j 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 $W = V\mu_j - c_j$, so $\mathbb{E}[W] = Vp_j - c_j$ and, since transfers are internal, $W = U^{\text{buyer}} + \sum_{i=1}^N U_i^{\text{seller}}$.

Reports. At the implementation level, seller i submits $\hat{\theta}_i = (\hat{p}_i, \hat{c}_i)$, inducing the score $\hat{s}_i = V\hat{p}_i - \hat{c}_i$. Since allocation and transfer depend on the report only through $\hat{s}_i$, reports with the same score are outcome-equivalent, so we analyze incentives in the reduced-form score space. Under Dual Error, seller i's strategic type is the effective surplus $\bar{T}_i = Vg_i - c_i$. Theorem 4.2 shows that any report satisfying $V\hat{p}_i - \hat{c}_i = \bar{T}_i$ is weakly dominant in induced outcomes; (g_i, c_i) is one canonical representative, so the guarantee applies to effective-surplus implementation rather than coordinate-wise identification.

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 evalu-

[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.

ated performance, and H charges the winner for the externality imposed on others, aligning the best response with surplus-based truthful competition (see Theorem 4.2 for the formal statement).

4 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.

4.1 MECHANISM GOALS

To ensure a reliable and sustainable routing market, we follow FTMD (Porter et al., 2008) and consider four standard desiderata: **(1) Dominant-Strategy Incentive Compatibility (DSIC):** Truth-telling maximizes each seller'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 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 (Porter et al., 2008). 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 benchmark property:

Proposition 4.1 (Optimality of the Error-Free Setting)**.** *In the error-free setting, the mechanism satisfies all four desirable properties: DSIC, IR, CR, and EE.*

This establishes the error-free benchmark, showing optimality under idealized conditions. Under Dual Error, exact welfare efficiency may be lost. Under the private-evaluation-belief structure introduced below, EA-RAM remains dominant-strategy incentive compatible in effective-surplus space and individually rational, admits sufficient conditions for CR, and enjoys a bounded welfare loss.

4.2 STRATEGIC PROPERTIES

Let H denote the runner-up score and let $\mathcal{I}_i$ denote seller i's information at bidding time. We adopt the following private-evaluation-belief structure: conditional on $\mathcal{I}_i$, competitors' reports do not alter seller i's belief about the buyer's evaluation of its own output, i.e., $\mathbb{E}_i[\tilde{\mu}_i \mid H, \mathcal{I}_i] = \mathbb{E}_i[\tilde{\mu}_i \mid \mathcal{I}_i] = g_i$. Accordingly, for a unilateral report $\hat{s}_i$, seller i's interim subjective expected utility is $U_i^{\text{seller}}(\hat{s}_i) = \mathbb{E}_i[(\bar{T}_i - H)\mathbf{1}\{H \le \hat{s}_i\} \mid \mathcal{I}_i]$, where $\bar{T}_i = Vg_i - c_i$.

Theorem 4.2 (Dominant-Strategy Incentive Compatibility in Effective-Surplus Space)**.** *Under the private-evaluation-belief structure above, for seller i with effective-surplus type $\bar{T}_i = Vg_i - c_i$, reporting $\hat{s}_i = \bar{T}_i$ maximizes its subjective expected utility for every realization of the runner-up score H. Hence, truthful effective-surplus reporting is a weakly dominant strategy in the reduced-form mechanism.*

Thus, each seller is incentivized to align its reported score with its belief-adjusted effective surplus rather than with the unobserved ground-truth surplus. Under this report, seller i's interim expected utility is $\mathbb{E}_i[U_i^{\text{seller}}] = \mathbb{E}_i[(\bar{T}_i - H)\mathbf{1}\{H \le \bar{T}_i\}] \ge 0$.

Theorem 4.3 (Individual Rationality (IR))**.** *Under truthful effective-surplus reporting $\hat{s}_i = \bar{T}_i$, every seller satisfies interim IR under its subjective belief: $\mathbb{E}_i[U_i^{\text{seller}} \mid \mathcal{I}_i] \ge 0$.*

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

Robustness to correlated beliefs. The private-evaluation-belief condition above is exact and may be violated when competitors' scores reveal residual information about the common task. We therefore quantify how the incentive guarantee degrades under bounded dependence.

Proposition 4.4 (Approximate IC under Correlated Evaluation Beliefs)**.** *Suppose seller i's conditional evaluation belief satisfies $|\mathbb{E}_i[\tilde{\mu}_i \mid H, \mathcal{I}_i] - g_i| \le \delta_i$ for every runner-up score H in its support.*

Then reporting $\hat{s}_i = \bar{T}_i$ is $V\delta_i$-approximately dominant in the reduced-form mechanism: for every alternative report and every H, the conditional expected utility gain from deviating is at most $V\delta_i$. Exact DSIC is recovered when $\delta_i = 0$.

Thus, residual information contained in competitors' scores does not immediately destroy incentive alignment: its maximal effect scales linearly with the induced belief deviation δ_i.

Theorem 4.5 (Center Rationality (CR))**.** *Each of the following sufficient conditions guarantees CR: (A) $\mathbb{E}[H] \geq V\Delta_{gate}$, where $\Delta_{gate} = L_\sigma\sqrt{b_{post} + a_{post}^2}$; (B) $\Delta_{cons} = \mathbb{E}\big[p_{(1)} - h_{(1)}\big] \geq 0$.*

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 4.6 (Comparative Statics: Ability and Difficulty)**.** *Holding fixed the distribution of H, seller i's interim subjective expected utility is weakly increasing in model ability m_i and weakly decreasing in 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.

4.3 Welfare Analysis

For welfare accounting, let the expected welfare under a specified selected seller i be $\mathbb{E}[W_i] = Vp_i - c_i$. We extend the seller set to $\bar{\mathcal{I}} = \mathcal{I} \cup \{0\}$, where seller 0 represents the null allocation with $p_0 = g_0 = h_0 = c_0 = 0$. We define the *true optimal allocation index* under the error-free setting as $i^\star \in \arg\max_{j\in\bar{\mathcal{I}}}\{Vp_j - c_j\}$, and the *error-aware allocation index* as $i^\dagger \in \arg\max_{j\in\bar{\mathcal{I}}}\{Vg_j - c_j\}$.

Proposition 4.7 (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 allocation index $i^\dagger$ 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 4.8 (Welfare-loss bound)**.** *Let $i^\star \in \arg\max_{i\in\bar{\mathcal{I}}}\{Vp_i - c_i\}$ and $i^\dagger \in \arg\max_{i\in\bar{\mathcal{I}}}\{Vg_i - c_i\}$. Then the expected welfare loss satisfies $0 \leq \mathbb{E}[W_{i^\star}] - \mathbb{E}[W_{i^\dagger}] \leq 2VL_\sigma(M_{\text{post}} + M_{\text{ante}})$.*

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

4.4 Structural Insights

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

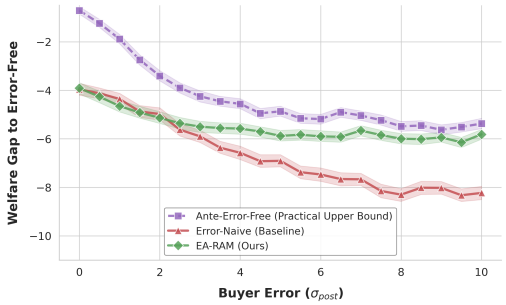
Proposition 4.9 (Opposite-Signed Error Compensation)**.** *If $(g_i - h_i)(h_i - p_i) \leq 0$, for $i \in \{i^\star, i^\dagger\}$, then the welfare-loss bound tightens to $2VL_\sigma \max\{M_{\text{ante}}, M_{\text{post}}\}$.*

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

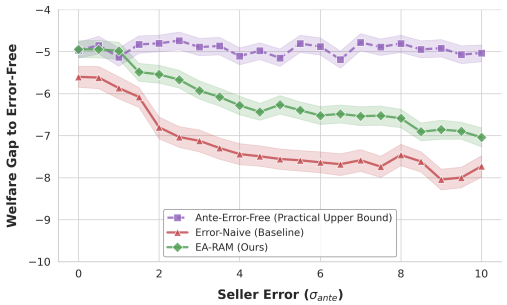
Proposition 4.10 (Saturation Robustness)**.** *In addition to the baseline link-function assumptions, suppose σ has vanishing tails, i.e., $\lim_{|x|\to\infty}\sigma'(x) = 0$. For any error ϵ with $\mathbb{E}[\epsilon^2] < \infty$, the deviation $\Delta(\phi_i) = |\mathbb{E}[\sigma(\phi_i + \epsilon)] - \sigma(\phi_i)|$ 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 4.11 (Noise-Induced Flattening)**.** *Let η be an independent noise term with $\mathbb{E}[|\eta|] < \infty$. Define the perturbed belief maps by convolution: $\tilde{g}_i(\phi) = \mathbb{E}[g_i(\phi + \eta)]$ and $\tilde{h}_i(\phi) = \mathbb{E}[h_i(\phi + \eta)]$. If g_i and h_i are continuously differentiable with bounded derivatives, then $\sup_\phi |\frac{\partial\tilde{g}_i}{\partial\phi}(\phi)| \leq$*



(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: EA-RAM remains stable under evaluator and prediction noise, avoiding the misallocation caused by naive reporting. Shaded regions show 95% confidence intervals.

$\sup_\phi |\frac{\partial g_i}{\partial \phi}(\phi)|$ *and* $\sup_\phi |\frac{\partial \tilde{h}_i}{\partial \phi}(\phi)| \leq \sup_\phi |\frac{\partial h_i}{\partial \phi}(\phi)|$. *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, reducing local sensitivity to strategic perturbations.

5 Experiments

5.1 Simulation Experiments

Table 2: Simulation Settings.

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

Experimental Setup. In this section, we run simulation experiments that explicitly control the magnitude of the Dual Error to evaluate EA-RAM's robustness. We simulate $N = 5$ heterogeneous sellers competing for one task with V=20 and d=1.5 (Table 2). 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 the buyer's acceptance probability h_i and reports accordingly; (2) **Error-Naive (Baseline):** sellers that form predictions without accounting for evaluator-side distortion; and (3) **EA-RAM (Ours):** sellers report according to their subjective acceptance belief g_i.

Error Robustness Analysis Results. 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 mitigates misallocation and makes the degradation milder.

5.2 Real-World Experiments

5.2.1 Experimental Setup

Benchmark and baselines. In this section, we benchmark EA-RAM against centralized baselines on real-world benchmarks. We evaluate routing policies on RouterBench (Hu et al., 2024), 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, coding, math, and knowledge-intensive understanding, and split them into train/test with a 70/30 ratio. We compare EA-RAM against a diverse set of strong baselines: EmbedLLM (Zhuang et al., 2025), IRT-Router (Song et al., 2025), RouteLLM (Ong et al., 2025), FrugalGPT (Chen et al., 2024), and Cascade Routing (Dekoninck et al., 2025). They cover a wide range of architectures, enabling a comprehensive comparison.

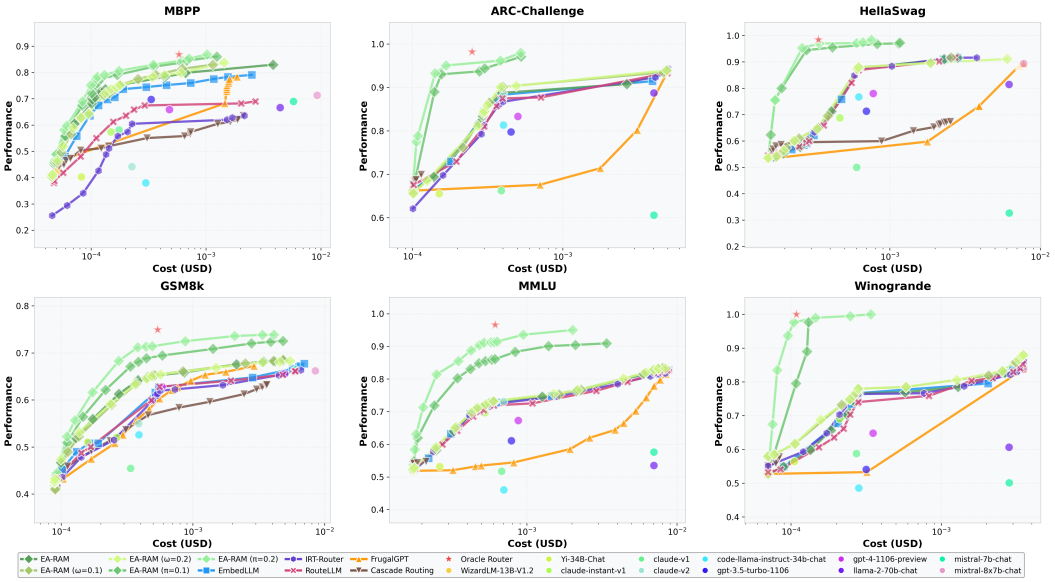


Figure 6: Cost–performance Pareto frontiers on real-world benchmarks. As π or ω increases, EA-RAM shifts the frontier upward and leftward, indicating a more favorable cost–quality trade-off than the centralized baselines.

Implementation details. EA-RAM trains two types of predictive modules. Each LLM has a seller-side ex-ante predictor that estimates query-level correctness from the query embedding, serving as an ex-ante proxy for the evaluator's acceptance probability. The buyer maintains an ex-post evaluator that estimates answer correctness from concatenated query and answer embeddings. Thus, both modules target answer correctness under different information sets: the seller predicts it ex ante from the query alone, whereas the buyer evaluates it ex post after observing the generated answer. Both modules are two-layer MLPs. The embedding model is `all-MiniLM-L6-v2` (Sentence-Transformers). More details are provided in Appendix A.3.

5.2.2 Routing Pareto Analysis

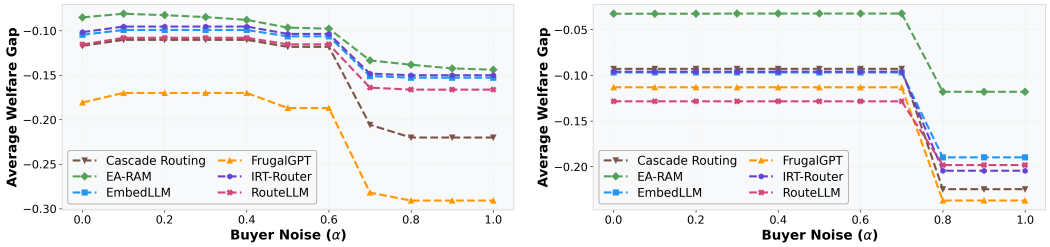
Table 3: Quantitative comparison (AIQ↑). EA-RAM yields the best overall performance.

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

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 $\omega \in [0,1]$. Figure 6 and Table 3 show that EA-RAM achieves a consistently favorable cost–performance trade-off. Table 3 reports *Average Improvement in Quality* (AIQ; Appendix A.2), which averages Pareto-frontier quality over a shared cost range. In the base setting, **EA-RAM achieves the highest AIQ** across all six datasets. **More informative seller-side signals further improve the frontier**: oracle information provides an upper-bound sensitivity study, while retrieval-based local information yields consistent gains over the base setting.

5.2.3 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 score $= \alpha \cdot \text{score}_{\text{judge}} + (1-\alpha) \cdot$



(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.

$\text{score}_{\text{groundtruth}}$, 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.

5.2.4 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.

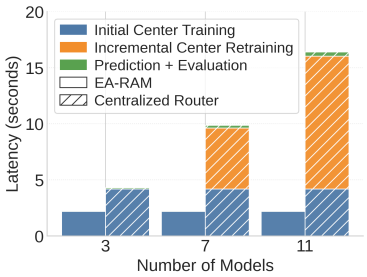


Figure 8: Center-side latency.

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.

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 4, 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.**

Table 4: Extra reverse-auction communication overhead on MBPP.

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

6 Conclusion

We presented EA-RAM, an error-aware reverse-auction mechanism for LLM routing that shifts capability estimation from a centralized router to providers while retaining a model-agnostic evaluator. Under Dual Error and the private-evaluation-belief structure, EA-RAM is DSIC in effective-surplus space, satisfies IR, admits sufficient conditions for CR, and enjoys bounded welfare loss. We further establish robustness properties including error compensation, saturation, and noise-induced smoothing. Experiments show robust cost–performance trade-offs and gains from provider-side local information, supporting market-based routing for heterogeneous LLM systems.

AI USE STATEMENT

In this work, we used LLMs to assist with code generation and language polishing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provided a detailed description and implementation of our method in Section 5.1, Section 5.2.1, and Appendix A. Code is provided in anonymous.4open.science/r/EARAM-8E0D.

REFERENCES

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.

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.

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, 2026a.

Haolong Chen, Liang Zhang, Zhuo Li, Lei Xue, and Guanrxu Zhu. Erskill: Evolving for skill-guided adaptive memory retrieval. *arXiv preprint arXiv:2608.12720*, 2026b.

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

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

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

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

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. URL https://openreview.net/forum?id=AAl89VNNy1.

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*, pp. 713–722, 2024.

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*, pp. 144–155, 2024.

Yijia Fan, Jusheng Zhang, Kaitong Cai, Jing Yang, Chengpei Tang, Jian Wang, and Keze Wang. Cost-effective communication: An auction-based method for language agent interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29412–29420, 2026.

Theodore Groves. Incentives in teams. *Econometrica: Journal of the Econometric Society*, pp. 617–631, 1973.

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.

Sama Habibi and Ozgur Ercetin. Edge-llm inference with cost-aware layer allocation and adaptive scheduling. *IEEE Access*, 13:131614–131637, 2025.

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.

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

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. Routerbench: A benchmark for multi-LLM routing system. In *Agentic Markets Workshop at ICML 2024*, 2024. URL https://openreview.net/forum?id=IVXmV8Uxwh.

Qizhe Li, Haolong Chen, Shan Dai, Zhuo Li, Zhiwei Hu, Xuan Li, Guangxu Zhu, and Qingjiang Shi. Qoereasoner: An agentic reasoning framework for automated and explainable qoe diagnosis in rans. *arXiv preprint arXiv:2606.01925*, 2026.

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

Tianyi Niu, Justin Chen, Genta Indra Winata, Shi-Xiong Zhang, Supriyo Chakraborty, Sambit Sahu, Yue Zhang, Elias Stengel-Eskin, and Mohit Bansal. Routing with generated data: Annotation-free llm skill estimation and expert selection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32441–32466, 2026.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

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

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

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

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.

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

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 (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15629–15644, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.761. URL https://aclanthology.org/2025.acl-long.761/.

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

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.

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.

Minrui Xu, Dusit Niyato, Hongliang Zhang, Jiawen Kang, Zehui Xiong, Shiwen Mao, and Zhu Han. Cached model-as-a-resource: Provisioning large language model agents for edge intelligence in space-air-ground integrated networks. *IEEE Transactions on Networking*, 2026.

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*, pp. 1–8, 2013.

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 (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Haozhen Zhang, Tao Feng, and Jiaxuan You. Router-r1: Teaching llms multi-round routing and aggregation via reinforcement learning. *Advances in Neural Information Processing Systems*, 38: 141233–141265, 2026a.

Yiqun Zhang, Hao Li, Zihan Wang, Shi Feng, Xiaocui Yang, Daling Wang, Bo Zhang, Lei Bai, and Shuyue Hu. Mtrouter: Cost-aware multi-turn llm routing with history–model joint embeddings. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44206–44226, 2026b.

Zheyuan Zhang, Kaiwen Shi, Zhengqing Yuan, Zehong Wang, Tianyi Ma, Keerthiram Murugesan, Vincent Galassi, Chuxu Zhang, and Yanfang Ye. Agentrouter: A knowledge-graph-guided llm router for collaborative multi-agent question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 788–809, 2026c.

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

A 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.

We use queries from six benchmarks covering reasoning (HellaSwag (Zellers et al., 2019), Winogrande (Sakaguchi et al., 2021), ARC-Challenge (Clark et al., 2018)), coding (MBPP (Austin et al., 2021)), math (GSM8k (Cobbe et al., 2021)), and knowledge-intensive understanding (MMLU (Hendrycks et al., 2021)).

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 5.

Table 5: 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 4.1)**.** *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 $\mathbb{E}[r_i] = p_i V - H$ and

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

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

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

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 $\mathbb{E}[U_i^{\text{seller}}] = s_i - H \geq 0$ because allocation requires $\hat{s}_i = s_i \geq H$. If i loses then $\mathbb{E}[U_i^{\text{seller}}] = 0$. Hence expected utility is never negative.

CR: In the error-free setting, $h_j = p_j$, so $\mathbb{E}[r_j] = Vp_j - H$. Therefore, $\mathbb{E}[U^{\text{buyer}}] = Vp_j - \mathbb{E}[r_j] = H \geq 0$. If the null allocation is selected, the buyer obtains zero utility.

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 (Dominant-Strategy Incentive Compatibility in Effective-Surplus Space. Restatement of Theorem 4.2)**.** *Under the private-evaluation-belief structure above, for seller i with effective-surplus type $\bar{T}_i = Vg_i - c_i$, reporting $\hat{s}_i = \bar{T}_i$ maximizes its subjective expected utility for every realization of the runner-up score H. Hence, truthful effective-surplus reporting is a weakly dominant strategy in the reduced-form mechanism.*

Proof. Under the private-evaluation-belief structure, conditional on seller i's information and a realized runner-up score H, its subjective expected payoff is $\bar{T}_i - H$ if it wins and 0 otherwise. Reporting $\hat{s}_i = \bar{T}_i$ therefore wins exactly when winning yields non-negative subjective expected utility. Any $\hat{s}_i < \bar{T}_i$ may forgo a profitable allocation, whereas any $\hat{s}_i > \bar{T}_i$ may induce an allocation with negative subjective expected utility. Since this argument holds for every H, $\hat{s}_i = \bar{T}_i$ is a weakly dominant strategy. □

Theorem C.3 (Individual Rationality (IR). Restatement of Theorem 4.3)**.** *Under truthful effective-surplus reporting $\hat{s}_i = \bar{T}_i$, every seller satisfies interim IR under its subjective belief: $\mathbb{E}_i[U_i^{\text{seller}} \mid \mathcal{I}_i] \geq 0$.*

Proof. Under truthful effective-surplus reporting, seller i's interim subjective expected utility is $\mathbb{E}_i[(\bar{T}_i - H)\mathbf{1}\{H \leq \bar{T}_i\} \mid \mathcal{I}_i]$. Since $\bar{T}_i - H \geq 0$ whenever $H \leq \bar{T}_i$, this expectation is non-negative. Hence, $\mathbb{E}_i[U_i^{\text{seller}} \mid \mathcal{I}_i] \geq 0$. □

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

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

Step 1: Bounding the deviation via the Lipschitz constant. Since σ is differentiable and satisfies $|\sigma'(z)| \leq L_\sigma$ for all z, the Mean Value Theorem implies that σ is L_σ-Lipschitz. Hence, for $a = x+\epsilon$ and $b = x$,

$$|\sigma(x+\epsilon) - \sigma(x)| \leq L_\sigma |\epsilon|. \tag{4}$$

Taking expectations and applying Jensen's inequality gives

$$\begin{aligned} \big|\mathbb{E}[\sigma(x+\epsilon)] - \sigma(x)\big| &= \big|\mathbb{E}[\sigma(x+\epsilon) - \sigma(x)]\big| \\ &\leq \mathbb{E}\big[|\sigma(x+\epsilon) - \sigma(x)|\big] \\ &\leq L_\sigma \mathbb{E}[|\epsilon|], \end{aligned} \tag{5}$$

where the last inequality follows from Eq. equation 4.

Step 2: Bounding the first absolute moment. By Jensen's inequality,

$$\mathbb{E}[|\epsilon|] \leq \sqrt{\mathbb{E}[\epsilon^2]}. \tag{6}$$

Using $\mathbb{E}[\epsilon^2] = \mathrm{Var}(\epsilon) + (\mathbb{E}[\epsilon])^2$ in Eq. equation 6, we obtain

$$\mathbb{E}[|\epsilon|] \leq \sqrt{\mathrm{Var}(\epsilon) + (\mathbb{E}[\epsilon])^2}. \tag{7}$$

Combining Eq. equation 5 and Eq. equation 7 yields

$$\big|\mathbb{E}[\sigma(x+\epsilon)] - \sigma(x)\big| \leq L_\sigma \mathbb{E}[|\epsilon|] \leq L_\sigma \sqrt{\mathrm{Var}(\epsilon) + (\mathbb{E}[\epsilon])^2}, \tag{8}$$

which proves the claim. □

Proposition C.5 (Approximate IC under Correlated Evaluation Beliefs. Restatement of Proposition 4.4)**.** *Suppose seller i's conditional evaluation belief satisfies $|\mathbb{E}_i[\tilde{\mu}_i \mid H, \mathcal{I}_i] - g_i| \le \delta_i$ for every runner-up score H in its support. Then reporting $\hat{s}_i = \bar{T}_i$ is $V\delta_i$-approximately dominant in the reduced-form mechanism: for every alternative report and every H, the conditional expected utility gain from deviating is at most $V\delta_i$. Exact DSIC is recovered when $\delta_i = 0$.*

Proof. Let $q_i(H) = \mathbb{E}_i[\tilde{\mu}_i \mid H, \mathcal{I}_i]$ and define $\xi_i(H) = V(q_i(H) - g_i)$, so that $|\xi_i(H)| \le V\delta_i$. Conditional on H, seller i's expected payoff from winning is $\bar{T}_i - H + \xi_i(H)$.

Reporting $\hat{s}_i = \bar{T}_i$ selects winning exactly when $H \le \bar{T}_i$. If an alternative report induces the same allocation, the utilities coincide. If truthful reporting wins while the alternative loses, the alternative can improve utility only when the winning payoff is negative; since $\bar{T}_i - H \ge 0$, this improvement is at most $V\delta_i$. Conversely, if truthful reporting loses while the alternative wins, then $\bar{T}_i - H < 0$, so the alternative's gain is again at most $V\delta_i$. Therefore truthful effective-surplus reporting is $V\delta_i$-approximately dominant. □

Theorem C.6 (Center Rationality (CR). Restatement of Theorem 4.5)**.** *Each of the following sufficient conditions guarantees CR: (A) $\mathbb{E}[H] \ge V\Delta_{gate}$, where $\Delta_{gate} = L_\sigma\sqrt{b_{post} + a_{post}^2}$; (B) $\Delta_{cons} = \mathbb{E}\big[p_{(1)} - h_{(1)}\big] \ge 0$.*

Proof. If the null allocation is selected, the buyer obtains zero utility and CR holds trivially. We therefore consider a non-null selected seller, indexed by (1), with runner-up score H.

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

$$r_{(1)} = V\tilde{\mu}_{(1)} - H, \tag{9}$$

the buyer's realized utility is

$$U^{\text{buyer}} = V\mu_{(1)} - r_{(1)} = H + V\big(\mu_{(1)} - \tilde{\mu}_{(1)}\big). \tag{10}$$

Taking expectations gives

$$\mathbb{E}[U^{\text{buyer}}] = \mathbb{E}[H] + V\mathbb{E}\big[p_{(1)} - h_{(1)}\big]. \tag{11}$$

(A) By the definition $h_i = \mathbb{E}_{\varepsilon_{\text{post}}}[\sigma(\phi_i + \varepsilon_{\text{post}})]$ and Lemma C.4, every seller i satisfies

$$|h_i - p_i| = \big|\mathbb{E}_{\varepsilon_{\text{post}}}[\sigma(\phi_i + \varepsilon_{\text{post}})] - \sigma(\phi_i)\big| \le L_\sigma\sqrt{b_{\text{post}} + a_{\text{post}}^2} = \Delta_{\text{gate}}. \tag{12}$$

Since this bound holds uniformly for every seller, it also holds for the selected seller (1). Hence,

$$\mathbb{E}\big[p_{(1)} - h_{(1)}\big] \ge -\Delta_{\text{gate}}. \tag{13}$$

Substituting equation 13 into equation 11 yields

$$\mathbb{E}[U^{\text{buyer}}] \ge \mathbb{E}[H] - V\Delta_{\text{gate}}. \tag{14}$$

Therefore, if $\mathbb{E}[H] \ge V\Delta_{\text{gate}}$, then $\mathbb{E}[U^{\text{buyer}}] \ge 0$, establishing CR.

(B) By the definition $\Delta_{\text{cons}} = \mathbb{E}[p_{(1)} - h_{(1)}]$, Eq. equation 11 gives

$$\mathbb{E}[U^{\text{buyer}}] = \mathbb{E}[H] + V\Delta_{\text{cons}}. \tag{15}$$

Because the dummy seller has score 0, the runner-up score satisfies $H \ge 0$, and hence $\mathbb{E}[H] \ge 0$. Therefore, whenever $\Delta_{\text{cons}} \ge 0$, Eq. equation 15 implies $\mathbb{E}[U^{\text{buyer}}] \ge 0$, establishing CR. □

Proposition C.7 (Comparative Statics: Ability and Difficulty. Restatement of Proposition 4.6)**.** *Holding fixed the distribution of H, seller i's interim subjective expected utility is weakly increasing in model ability m_i and weakly decreasing in task difficulty d.*

Proof. Holding the distribution of H fixed, seller i's interim subjective utility under truthful effective-surplus reporting is $\mathbb{E}_i[(\bar{T}_i - H)_+ \mid \mathcal{I}_i]$, where $\bar{T}_i = Vg_i - c_i$. Since σ is strictly increasing, g_i is non-decreasing in ϕ_i. Moreover, ϕ_i is non-decreasing in m_i and non-increasing in d. Hence, $\bar{T}_i$ is weakly increasing in m_i and weakly decreasing in d. Because $(T - H)_+$ is non-decreasing in T for every H, the claimed monotonicity follows. □

C.3 WELFARE ANALYSIS

Proposition C.8 (Economic Efficiency (EE). Restatement of Proposition 4.7)**.** *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 allocation index $i^\dagger$ is welfare-optimal.*

Proof. Define $W_0 = 0$. We adopt the convention $p_0 = g_0 = h_0 = c_0 = 0$ for the dummy seller. Let $i^\star$ denote an error-free welfare-maximizing seller. If seller i is selected, then the realized welfare is

$$W_i = V\mu_i - c_i. \tag{16}$$

Taking the expectation gives

$$\mathbb{E}[W_i] = V\mathbb{E}[\mu_i] - c_i = Vp_i - c_i. \tag{17}$$

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

$$\mathbb{E}[W_{i^\dagger}] = Vp_{i^\dagger} - c_{i^\dagger}, \tag{18}$$

whereas the error-free benchmark welfare is

$$\mathbb{E}[W_{i^\star}] = \max_{i\in\bar{\mathcal{I}}}\{Vp_i - c_i\}. \tag{19}$$

Therefore,

$$\mathbb{E}[W_{i^\dagger}] = \mathbb{E}[W_{i^\star}] \iff Vp_{i^\dagger} - c_{i^\dagger} = \max_{i\in\bar{\mathcal{I}}}\{Vp_i - c_i\}. \tag{20}$$

This proves the claim. □

Theorem C.9 (Welfare-loss bound. Restatement of Theorem 4.8)**.** *Let $\bar{\mathcal{I}} = \mathcal{I} \cup \{0\}$ include the dummy seller representing the null allocation, with $p_0 = g_0 = h_0 = c_0 = 0$. Let $i^\star \in \arg\max_{i\in\bar{\mathcal{I}}}\{Vp_i - c_i\}$ and $i^\dagger \in \arg\max_{i\in\bar{\mathcal{I}}}\{Vg_i - c_i\}$. Then the expected welfare loss satisfies $0 \le \mathbb{E}[W_{i^\star}] - \mathbb{E}[W_{i^\dagger}] \le 2VL_\sigma(M_{\text{post}} + M_{\text{ante}})$.*

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

$$\begin{aligned}\mathbb{E}[W_{i^\star}] - \mathbb{E}[W_{i^\dagger}] &= \big(\mathbb{E}[W_{i^\star}] - \bar{T}_{i^\star}\big) + \big(\bar{T}_{i^\star} - \bar{T}_{i^\dagger}\big) + \big(\bar{T}_{i^\dagger} - \mathbb{E}[W_{i^\dagger}]\big)\\ &\le \big(\mathbb{E}[W_{i^\star}] - \bar{T}_{i^\star}\big) + \big(\bar{T}_{i^\dagger} - \mathbb{E}[W_{i^\dagger}]\big).\end{aligned} \tag{21}$$

Moreover, since $i^\star$ maximizes the true expected welfare over $\bar{\mathcal{I}}$, the welfare gap is non-negative. Therefore,

$$0 \le \mathbb{E}[W_{i^\star}] - \mathbb{E}[W_{i^\dagger}] \le V\Big(|g_{i^\star} - p_{i^\star}| + |g_{i^\dagger} - p_{i^\dagger}|\Big). \tag{22}$$

Step 2: Bound the two error channels. By the definitions of the buyer's acceptance probability and the seller's subjective acceptance probability, $h_i = \mathbb{E}_{\varepsilon_{\text{post}}}[\sigma(\phi_i + \varepsilon_{\text{post}})]$ and $g_i = \mathbb{E}_{\varepsilon_{\text{post}},\varepsilon_{\text{ante},i}}[\sigma(\phi_i + \varepsilon_{\text{post}} + \varepsilon_{\text{ante},i})]$. Hence,

$$|g_i - p_i| \le |g_i - h_i| + |h_i - p_i|. \tag{23}$$

Applying Lemma C.4 to the evaluator-side error gives

$$|h_i - p_i| \le L_\sigma\sqrt{b_{\text{post}} + a_{\text{post}}^2} = L_\sigma M_{\text{post}}. \tag{24}$$

Similarly, conditioning on $\varepsilon_{\text{post}}$ and applying the same lemma to the seller-side belief error yields

$$|g_i - h_i| \le L_\sigma\sqrt{b_{\text{ante},i} + a_{\text{ante},i}^2} \le L_\sigma M_{\text{ante}}. \tag{25}$$

Combining equation 23–equation 25,

$$|g_i - p_i| \le L_\sigma\big(M_{\text{post}} + M_{\text{ante}}\big) \qquad \text{for all } i \in \bar{\mathcal{I}}. \tag{26}$$

Step 3: Combine. Applying equation 26 to $i^\star$ and $i^\dagger$ in equation 22 gives

$$0 \le \mathbb{E}[W_{i^\star}] - \mathbb{E}[W_{i^\dagger}] \le 2VL_\sigma\big(M_{\text{post}} + M_{\text{ante}}\big), \tag{27}$$

which completes the proof. □

C.4 Structural Insights

Proposition C.10 (Opposite-Signed Error Compensation. Restatement of Proposition 4.9)**.** *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 $i^\star \in \arg\max_{j \in \bar{\mathcal{I}}}\{Vp_j - c_j\}$ denote the welfare-maximizing allocation under the error-free setting, and let $i^\dagger \in \arg\max_{j \in \bar{\mathcal{I}}}\{Vg_j - c_j\}$ denote the allocation selected by the error-involved mechanism. We start from the welfare-gap decomposition established in Theorem 4.8:

$$0 \le \mathbb{E}[W_{i^\star}] - \mathbb{E}[W_{i^\dagger}] \le V\Big(|g_{i^\star} - p_{i^\star}| + |g_{i^\dagger} - p_{i^\dagger}|\Big). \tag{28}$$

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:

$$|g_i - p_i| = |a_i + b_i| \le \max\{|a_i|, |b_i|\}. \tag{29}$$

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

$$|g_{i^\star} - p_{i^\star}| + |g_{i^\dagger} - p_{i^\dagger}| \;\le\; \max\{|a_{i^\star}|, |b_{i^\star}|\} + \max\{|a_{i^\dagger}|, |b_{i^\dagger}|\}. \tag{30}$$

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

$$|g_i - h_i| \le L_\sigma\, M_{\text{ante}}, \tag{31}$$

$$|h_i - p_i| \le L_\sigma\, M_{\text{post}}, \tag{32}$$

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

$$\sup_j |g_j - h_j| \le L_\sigma M_{\text{ante}}, \quad \sup_j |h_j - p_j| \le L_\sigma M_{\text{post}}. \tag{33}$$

Combining equation 28–equation 33, we obtain

$$0 \;\le\; \mathbb{E}[W_{i^\star}] - \mathbb{E}[W_{i^\dagger}] \;\le\; 2VL_\sigma\, \max\big\{M_{\text{ante}},\, M_{\text{post}}\big\}. \tag{34}$$

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

$$2VL_\sigma\, \max\big\{M_{\text{ante}},\, M_{\text{post}}\big\} \;<\; 2VL_\sigma\,(M_{\text{post}} + M_{\text{ante}}), \tag{35}$$

which completes the proof. □

Proposition C.11 (Saturation Robustness. Restatement of Proposition 4.10)**.** $\lim_{|x|\to\infty} \sigma'(x) = 0$. *For any error* ϵ *with* $\mathbb{E}[\epsilon^2] < \infty$*, the deviation* $\Delta(\phi_i) = |\mathbb{E}[\sigma(\phi_i + \epsilon)] - \sigma(\phi_i)|$ *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,

$$\sigma(\phi_i + \epsilon) - \sigma(\phi_i) = \epsilon \int_0^1 \sigma'(\phi_i + t\epsilon)\, dt. \tag{36}$$

Taking absolute values and expectations gives

$$\Delta(\phi_i) \le \mathbb{E}\left[|\epsilon| \int_0^1 |\sigma'(\phi_i + t\epsilon)|\, dt\right]. \tag{37}$$

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

$$|x| \ge R \implies |\sigma'(x)| \le \eta. \tag{38}$$

Define the event $A = \{|\epsilon| \le |\phi_i|/2\}$ and split the expectation in equation 37 over A and A^c.

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

$$|\phi_i + t\epsilon| \ge |\phi_i| - t|\epsilon| \ge |\phi_i| - \frac{|\phi_i|}{2} = \frac{|\phi_i|}{2}. \tag{39}$$

Hence, if $|\phi_i| \geq 2R$, then $|\phi_i + t\epsilon| \geq R$ and by equation 38 we have $|\sigma'(\phi_i + t\epsilon)| \leq \eta$. Therefore,

$$\mathbb{E}\left[|\epsilon| \int_0^1 |\sigma'(\phi_i + t\epsilon)|\, dt\, \mathbf{1}_A\right] \leq \eta\, \mathbb{E}[|\epsilon|]. \tag{40}$$

On A^c, we use boundedness of σ':

$$\mathbb{E}\left[|\epsilon| \int_0^1 |\sigma'(\phi_i + t\epsilon)|\, dt\, \mathbf{1}_{A^c}\right] \leq \|\sigma'\|_\infty\, \mathbb{E}[|\epsilon|\, \mathbf{1}\{|\epsilon| > |\phi_i|/2\}]\,. \tag{41}$$

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

$$\mathbb{E}[|\epsilon|\, \mathbf{1}\{|\epsilon| > t\}] \xrightarrow[t\to\infty]{} 0 \tag{42}$$

(e.g., by dominated convergence with the dominating integrable random variable $|\epsilon|$). Thus the right-hand side of equation 41 vanishes as $|\phi_i| \to \infty$.

Combining equation 37–equation 41 yields, for all sufficiently large $|\phi_i|$,

$$\Delta(\phi_i) \leq \eta\, \mathbb{E}[|\epsilon|] + \|\sigma'\|_\infty\, \mathbb{E}[|\epsilon|\, \mathbf{1}\{|\epsilon| > |\phi_i|/2\}]\,. \tag{43}$$

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

Proposition C.12 (Noise-Induced Flattening. Restatement of Proposition 4.11)**.** *Let η be an independent noise term with $\mathbb{E}[|\eta|] < \infty$. Define the perturbed belief maps by convolution: $\tilde{g}_i(\phi) = \mathbb{E}[g_i(\phi + \eta)]$ and $\tilde{h}_i(\phi) = \mathbb{E}[h_i(\phi + \eta)]$. If g_i and h_i are continuously differentiable with bounded derivatives, then $\sup_\phi |\frac{\partial \tilde{g}_i}{\partial \phi}(\phi)| \leq \sup_\phi |\frac{\partial g_i}{\partial \phi}(\phi)|$ and $\sup_\phi |\frac{\partial \tilde{h}_i}{\partial \phi}(\phi)| \leq \sup_\phi |\frac{\partial h_i}{\partial \phi}(\phi)|$. 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,

$$\tilde{g}_i(\phi) = \mathbb{E}[g_i(\phi + \eta)]. \tag{44}$$

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

$$\frac{\partial \tilde{g}_i}{\partial \phi}(\phi) = \mathbb{E}[g_i'(\phi + \eta)]\,. \tag{45}$$

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

$$|\frac{\partial \tilde{g}_i}{\partial \phi}(\phi)| = |\mathbb{E}[g_i'(\phi + \eta)]| \leq \mathbb{E}[|g_i'(\phi + \eta)|] \leq \sup_y |g_i'(y)|. \tag{46}$$

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

$$\sup_\phi |\frac{\partial \tilde{g}_i}{\partial \phi}(\phi)| \leq \sup_y |g_i'(y)|, \tag{47}$$

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 DSIC, 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.