

Cost-LLM: Efficient Multimodal Procurement-Cost Reasoning for E-Commerce Price Governance

Zhiyin Li*, Fei Huang*, Tengpeng Kong[†], Jie Tang, Shengjie Li

JD.com, Beijing, China

{lizhiyin.9, huangfei129, kongtengteng3, tangjie18, lishengjie1}@jd.com

Abstract

Estimation of product procurement cost is a key capability for price governance on large-scale e-commerce platforms: from images, structured attributes, comparable products, and pricing knowledge, a practical system must infer a reasonable cost to surface suspiciously inflated or potentially non-compliant listings. Directly invoking general-purpose VLMs is prohibitive at high QPS (queries per second), and their verbose rationales complicate auditing and rule integration. We present COST-LLM, a compact 7B multimodal cost-estimation model trained by a two-stage post-training pipeline: segment-normalized SFT on progressive pricing rationales, followed by GRPO with an asymmetric MAPE-based reward aligned with cost accuracy. On an in-house benchmark covering 105 fine-grained categories, COST-LLM reaches MAPE 29.47% and Acc@20 40.42%, yielding the best point estimates among the evaluated systems. Deployed as an auxiliary signal in a price-anomaly pipeline, it provides a 30.73% relative complementary-detection yield over the production baseline at an audited precision-matched operating point. Its locally deployed serving latency is 0.4–0.8 s per listing, compared with 3–8 s observed for external commercial LLM/VLM APIs. We share the design choices, training recipe, and deployment lessons.

1 Introduction

Price governance is a key component of risk control on large-scale e-commerce platforms: listings can exhibit suspiciously inflated prices due to abnormal pricing, misleading positioning, or other rule violations, and surfacing such cases requires estimating a reasonable *procurement cost* rather than just predicting the selling price. From prod-

uct images, structured attributes, comparable products, and pricing knowledge, the system must infer the underlying cost as an auxiliary signal for identifying overpriced or non-compliant listings.

This task is challenging for two reasons. Procurement cost is not directly observable from the listing price and must be inferred from heterogeneous multimodal evidence; meanwhile, industrial risk control demands stable output format, low serving cost, and high-throughput inference. Although large LLMs/VLMs with chain-of-thought prompting reason strongly (Wei et al., 2022; Zhang et al., 2024; DeepSeek-AI et al., 2025), serving them at high QPS (queries per second) is expensive and their long rationales can be redundant or inconsistent.

We propose COST-LLM, a compact 7B multimodal cost-estimation model built on Qwen2.5-VL-7B-Instruct. We compress long teacher rationales into minimal-sufficient progressive pricing rationales, apply segment-normalized SFT to balance format/rationale/answer supervision, and then run GRPO with an asymmetric MAPE-based reward aligned with cost accuracy. On an in-house benchmark and in live deployment, COST-LLM matches or beats much larger general-purpose foundation models on reported point estimates while providing complementary detections beyond an existing comparative-pricing baseline.

We position this work as an industry-track contribution: rather than pushing state-of-the-art on a generic benchmark, we describe what it takes to move a vertical reasoning capability from an offline metric to a deployed risk-control signal. We document concrete deployment-time choices—segment-normalized supervision so that long rationales do not dominate the loss, an asymmetric MAPE reward that suppresses the over-estimation bias caused by anchoring on transaction-priced comparables, an operating-point selection that matches the existing online baseline’s audited pre-

* Equal contribution.

[†] Corresponding author.

cision, and an on-premise compact-model serving stack that removes per-call API fees and avoids the network, queueing, and long-output overhead of external APIs—and report the lessons learned from auditing the resulting detections.

Our contributions are: (i) we formulate procurement-cost estimation as a multimodal reasoning problem for e-commerce risk control, distinct from selling-price prediction; (ii) we propose COST-LLM, trained with progressive rationale construction, segment-normalized SFT, and asymmetric MAPE-reward GRPO under a high-QPS serving budget; (iii) we conduct offline benchmarking and a live online evaluation, showing competitive accuracy against much larger models and a 30.73% relative complementary-detection yield over the production baseline at an audited precision-matched operating point; and (iv) we share deployment lessons—operating-point calibration, latency vs. external-API trade-offs, and rationale-answer consistency as a deployment-time diagnostic—useful for teams running compact reasoning LLMs in production.

2 Related Work

Multimodal reasoning and pricing. Chain-of-thought prompting (Wei et al., 2022) and its multimodal extensions (Zhang et al., 2024) improve reasoning, and recent VLMs such as Qwen2.5-VL (Bai et al., 2025) and InternVL (Chen et al., 2024) further strengthen visual-textual perception. LLP (Wang et al., 2025) studies LLM-based selling-price prediction with retrieval, SFT, and GRPO; in contrast, we target *procurement-cost* estimation for downstream price-governance decisions.

Post-training and efficient reasoning. RL-based post-training is effective when verifiable rewards exist (DeepSeek-AI et al., 2025; Kimi Team et al., 2025); we follow this paradigm with a MAPE-based reward. For deployment, long CoT improves hard reasoning but inflates token cost and may inject inconsistency, motivating token-budget, early-exit, and overthinking-mitigation techniques (Han et al., 2025; Sui et al., 2025; Yang et al., 2025). These observations directly motivate our compact, concise-rationale post-training for high-QPS risk control.

3 Task and Method

3.1 Procurement-Cost Estimation

For each product, we denote the multimodal input and reference procurement cost as $x_i = (I_i, A_i, C_i, K_i), p_i^* \in \mathbb{R}^+$, where I_i, A_i, C_i, K_i are the product image, structured attributes, comparable products, and pricing knowledge. The model outputs an estimated cost $\hat{p}$ with an optional rationale r : $(\hat{p}, r) = f_\theta(I, A, C, K)$. Unlike selling-price regression, the target is the underlying procurement cost, which depends on material, manufacturing, brand premium, channel factors, and inventory risk—thus requiring evidence-aware multimodal reasoning rather than purely numerical fitting.

Pricing knowledge construction. The pricing-knowledge component K is built from a category-conditioned cost knowledge base curated from platform pricing-governance accumulation and LLM-assisted enrichment. For each fine-grained category, the knowledge base enumerates the dominant cost drivers and their typical cost ranges—for instance, in the *phone case* category the core material is one of *ordinary TPU / liquid silicone / carbon fiber / etc.*, each with a distinct procurement-cost band; analogous fields cover manufacturing process, structural complexity, and brand-tier premium. Given a target listing, we retrieve the most relevant entries via category- and attribute-aware RAG and concatenate them into K . This turns otherwise tacit cost priors into explicit, retrievable evidence and gives the model an anchor for calibrated cost reasoning that purely visual features rarely provide.

3.2 Progressive Pricing-Rationale Construction

Since raw records contain only reference prices, we first synthesize long pricing rationales with a strong multimodal teacher $\mathcal{T}$:

$$r_i^{\text{cot}} = \mathcal{T}(x_i) = (s_{i,1}, s_{i,2}, \dots, s_{i,T_i}). \quad (1)$$

Here $s_{i,j}$ denotes the j -th reasoning segment of the i -th teacher rationale, and T_i is the number of segments in that rationale. To remove redundant reasoning, each rationale is decomposed into prefixes $r_i^{(t)} = (s_{i,1}, \dots, s_{i,t})$, and a frozen pricing model g_ϕ evaluates each prefix:

$$\hat{p}_i^{(t)} = g_\phi(x_i, r_i^{(t)}), \quad e_i^{(t)} = \frac{|\hat{p}_i^{(t)} - p_i^*|}{p_i^*}. \quad (2)$$

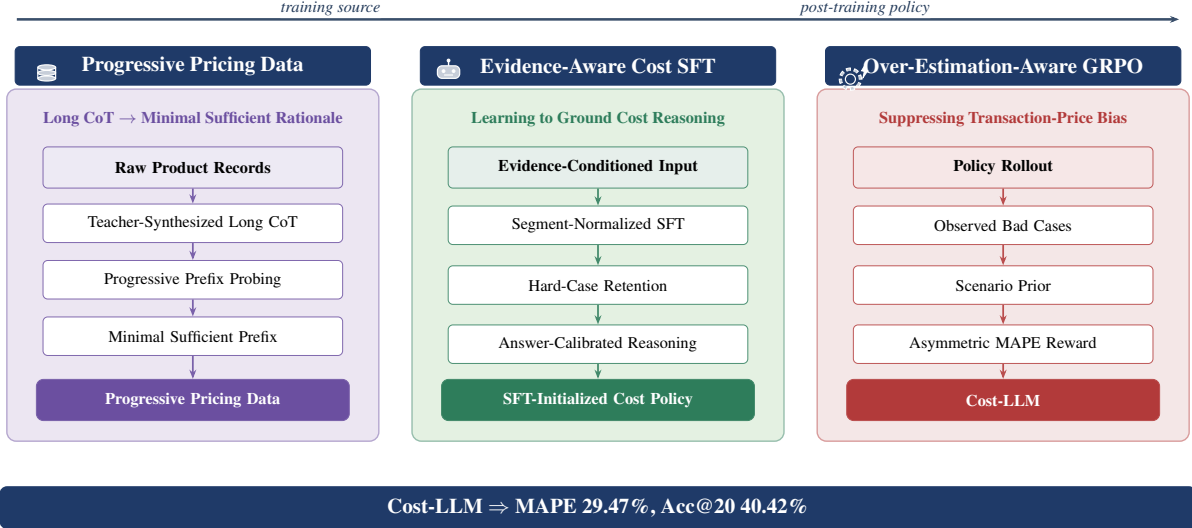


Figure 1: Overview of COST-LLM. The framework first constructs progressive pricing data by compressing long teacher rationales into minimal sufficient evidence chains. The resulting data are used for evidence-aware SFT with segment-normalized supervision, where multimodal product information, comparable products, and pricing knowledge are jointly grounded for procurement-cost estimation. Finally, MAPE-reward GRPO further aligns the model with procurement-cost accuracy.

Here $\hat{p}_i^{(t)}$ is the cost estimated from the length- t prefix and $e_i^{(t)}$ is its relative error against the reference cost p_i^* . We retain the shortest sufficient prefix

$$t_i^* = \min\{t \mid e_i^{(t)} \leq \epsilon\}, \quad (3)$$

yielding $\mathcal{D}_{\text{prog}} = \{(x_i, r_i^{(t_i^*)}, p_i^*)\}_{i=1}^N$. This provides adaptive-depth pricing supervision: simple products keep only decisive evidence, while complex products retain longer trajectories. For teacher rationales that admit no prefix satisfying $e_i^{(t)} \leq \epsilon$, we retain the complete long chain rather than discarding the sample, which avoids biasing the corpus toward easy, short-prefix cases (see Limitations). Such unmatched (hard) cases account for about 18% of the corpus. Because unmatched samples are kept rather than dropped, the corpus size is unchanged at N examples; progressive selection only shortens the supervised rationale, reducing the average rationale length from 1,722 to 1,246 characters ($\sim 27.6\%$ compression) while preserving all pricing-decisive evidence.

Implementation details. As illustrated in Figure 1 (left panel), the teacher $\mathcal{T}$ is Gemini-3-Pro Preview, which consumes (I, A, C, K) and emits a structured long-CoT rationale of up to $T_i \leq 1024$ tokens. The frozen prefix evaluator g_ϕ is the pretrained Qwen3-VL-32B used zero-shot, and we retain the shortest prefix whose relative error falls within the sufficiency threshold $\epsilon =$

0.20. Decoupling the high-capacity teacher from the lightweight prefix selector keeps long-CoT generation rich while making minimal-sufficient-trajectory extraction cheap and reproducible. Figure 2 summarizes this pipeline, and Figure 3 compares the rationale-length distributions before and after the progressive selection.

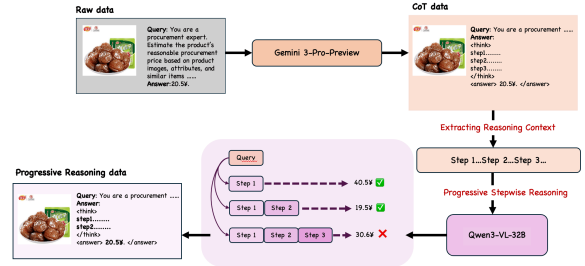


Figure 2: Pipeline for progressive pricing-rationale construction.

3.3 Cost-LLM Training Pipeline

We train COST-LLM from Qwen2.5-VL-7B-Instruct through two post-training stages on the progressive pricing-rationale corpus: SFT teaches an evidence-grounded reasoning paradigm, and RL further optimizes the final cost estimate toward the business metric. The pipeline balances two objectives: keeping the reasoning concise and structured, and calibrating the final numerical answer.

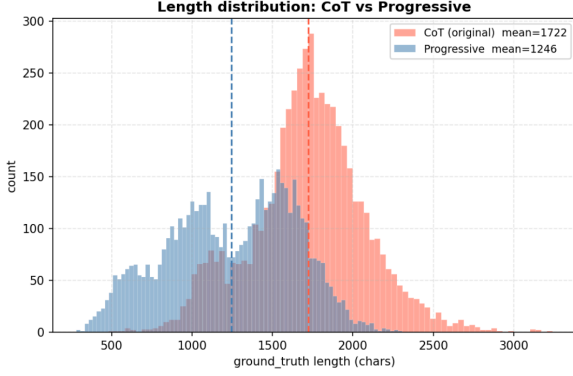


Figure 3: Rationale length distribution: long CoT vs. progressive rationale.

3.3.1 SFT: Segment-Normalized Cost Supervision

A standard SFT loss computes token-level negative log-likelihood over the entire response. In our task, however, the `<think>` segment is much longer than the `<answer>` segment, so a uniform token-level loss lets rationale tokens dominate the gradient while the final procurement-cost value receives limited supervision—undesirable for risk control where the answer field is the key signal.

To address this, we adopt a segment-normalized SFT objective. Given an input x_i and target response y_i , we partition the assistant-side tokens into three disjoint segments

$$\mathcal{M} = \mathcal{M}_{\text{fmt}} \cup \mathcal{M}_{\text{thk}} \cup \mathcal{M}_{\text{ans}}, \quad (4)$$

where $\mathcal{M}_{\text{fmt}}$ collects format tokens, $\mathcal{M}_{\text{thk}}$ the rationale tokens, and $\mathcal{M}_{\text{ans}}$ the final numerical-cost tokens. For each segment we compute the mean token-level cross-entropy

$$\mathcal{L}_s = \frac{1}{|\mathcal{M}_s|} \sum_{t \in \mathcal{M}_s} -\log \pi_{\theta}(y_{i,t} | x_i, y_{i,<t}), \quad (5)$$

with $s \in \{\text{fmt}, \text{thk}, \text{ans}\}$. The final SFT objective is a weighted average:

$$\mathcal{L}_{\text{SFT}} = \frac{\lambda_{\text{fmt}} \mathcal{L}_{\text{fmt}} + \lambda_{\text{thk}} \mathcal{L}_{\text{thk}} + \lambda_{\text{ans}} \mathcal{L}_{\text{ans}}}{\lambda_{\text{fmt}} + \lambda_{\text{thk}} + \lambda_{\text{ans}}}, \quad (6)$$

with $(\lambda_{\text{fmt}}, \lambda_{\text{thk}}, \lambda_{\text{ans}}) = (0.5, 0.5, 2.5)$. Segment-level normalization prevents long rationales from dominating purely by length, and the larger answer weight strengthens supervision on the final cost. To avoid biasing the progressive corpus toward easy cases, we retain the full teacher-generated chain for the $\approx 18\%$ of training samples whose rationales admit no sufficiently accurate short prefix.

We call this procedure *hard-case retention*; these retained samples are trained with the same objective in (6), without a separate SFT pass or a new loss.

3.3.2 RL: Asymmetric MAPE-Reward GRPO

After SFT, the model has a stable evidence-grounded output format, but supervised imitation alone cannot directly optimize the business-facing numerical metric. We therefore apply GRPO with an asymmetric MAPE-based reward. Given input $x = (I, A, C, K)$, the policy samples $y = (r, \hat{p}) \sim \pi_{\theta}(\cdot | x)$; we parse $\hat{p}$ from `<answer>`, compare it with the ground truth p^* , and compute the relative error $e(\hat{p}, p^*) = |\hat{p} - p^*|/p^*$. The reward is an asymmetric clipped MAPE-based score:

$$R_{\text{cost}}(y) = \text{clip}(0.5 - w(\hat{p}, p^*)e, -0.5, 0.5),$$

$$w(\hat{p}, p^*) = \begin{cases} 1.5, & \hat{p} > p^*, \\ 1, & \hat{p} \leq p^*. \end{cases} \quad (7)$$

A bad-case audit showed that the model tends to over-estimate procurement cost—comparable products may carry observed transaction prices, which include margin, platform premium, and logistics costs, so anchoring on them systematically biases the estimate upward. The $1.5\times$ penalty on $\hat{p} > p^*$ suppresses this bias while keeping the reward simple and verifiable.

For each input x , GRPO samples a group of G responses $\{y_j\}_{j=1}^G$ from $\pi_{\theta_{\text{old}}}$ with rewards $\{R_j\}_{j=1}^G$, and computes the group-normalized advantage $A_j = (R_j - \text{mean}(R))/(\text{std}(R) + \epsilon)$. Let $\bar{\rho}_j = \text{clip}(\rho_j, 1 - \epsilon_c, 1 + \epsilon_c)$ denote the clipped importance ratio. The policy is optimized with the GRPO surrogate

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_j[\min(\rho_j A_j, \bar{\rho}_j A_j)] + \beta D_{\text{KL}}(\pi_{\theta} \| \pi_{\text{ref}}). \quad (8)$$

with importance ratio $\rho_j = \pi_{\theta}(y_j | x)/\pi_{\theta_{\text{old}}}(y_j | x)$. π_{ref} is initialized from the SFT model; the KL term stabilizes training and prevents the policy from drifting away from the supervised cost-reasoning behavior. During RL we monitor MAPE, Acc@10/20, parse success, rationale-answer consistency, and response length for evaluation and model selection, while the optimization reward stays centered on procurement-cost accuracy.

On numerical regression via token-level decoding. A subtlety of using an LLM for procurement-cost estimation is that a continuous cost value is produced by decoding numeric tokens rather than by a dedicated regression head, which can introduce discretization and calibration effects. Two of our design choices directly narrow this gap. First, the segment-normalized SFT objective in (6) up-weights the answer segment ($\lambda_{\text{ans}} = 2.5$), so the numeric-token distribution is supervised more sharply than under a uniform token loss. Second, the asymmetric MAPE reward in (7) imposes a scale-aware, numerically consistent calibration signal on the decoded value rather than a purely token-level likelihood, pulling the emitted number toward the true cost and suppressing the systematic upward bias. We view alternative mechanisms—a dedicated numeric head, constrained/grammar-based decoding, or a hybrid regression-generation output—as complementary directions for future work.

4 Experiments

We evaluate COST-LLM on an in-house procurement-cost dataset of about 25K instances from 105 fine-grained categories (clothing, shoes, home appliances, accessories, daily commodities, etc.). Each instance carries a product image, title, category hierarchy, brand, structured attributes, comparable-product information, pricing knowledge, and a reference procurement cost aggregated from internal supplier-side records and platform pricing-governance audits (residual label noise from temporal drift is acknowledged as a limitation). We use a 0.75/0.10/0.15 split ($\sim 18.75\text{K}/2.5\text{K}/3.75\text{K}$). To prevent leakage, the split is performed at the *product-ID* level: once a product is assigned to test, its ID and all references to it (including where it appears as a *comparable* for other products) are removed from the comparable fields and retrieval candidates of every training sample; the pricing knowledge base K stores category-level aggregated priors rather than per-product ground-truth entries, and the teacher/prefix evaluator are only run on training-split products.

Starting from Qwen2.5-VL-7B-Instruct, the model samples $(r, \hat{p}) \sim \pi_{\theta}(\cdot \mid I, A, C, K)$ under the two-stage pipeline of Section 3.3: segment-normalized SFT on the progressive corpus with hard-case retention, then GRPO from the best

SFT checkpoint optimizing the reward in (7); checkpoints are selected by validation MAPE. We compare against API-based baselines—Qwen2.5-VL-72B, Doubao-Seed-1.6, Qwen3-235B-A22B, Qwen3-VL-235B, and Gemini-3-Flash-Preview—and a discriminative non-LLM regressor, PriceBERT, under an identical prompt and evidence inputs $x = (I, A, C, K)$. Detailed baseline prompts, input formatting, output-parsing rules, and token budgets are given in Appendix A.1 to ensure a fair, equally optimized comparison.

Evaluation metrics. We report MAPE $= \frac{100}{N} \sum_i |\hat{p}_i - p_i^*|/p_i^*$ and threshold accuracy $\text{Acc}@ \tau = \frac{1}{N} \sum_i \mathbb{I}[|\hat{p}_i - p_i^*|/p_i^* \leq \tau]$ with $\tau \in \{0.1, 0.2\}$; throughout, $\text{Acc}@10$ and $\text{Acc}@20$ denote $\text{Acc}@ \tau$ at $\tau = 0.1$ and $\tau = 0.2$. **Consistency** is the fraction of test instances whose $\hat{p}$ matches the last explicit price token in the rationale within $\pm 1\%$ tolerance, and **format validity** is the fraction of outputs conforming to the required `<think>...</think><answer>...</answer>` structure.

4.1 Results and Analysis

Table 1 reports the main comparison: COST-LLM obtains the best reported pricing accuracy among the evaluated systems (MAPE 29.47%, $\text{Acc}@10$ 19.73%, $\text{Acc}@20$ 40.42%) while keeping a structured format and full rationale-answer consistency. Paired-bootstrap tests in Appendix A.4 show that the 0.81-point MAPE gap from the closest API baseline, Doubao-Seed-1.6, is not statistically decisive, whereas the larger gaps from the remaining evaluated baselines are significant. We therefore do not claim a decisive accuracy advantage over Doubao; the deployment case instead combines competitive accuracy, full consistency, observed local latency of 0.4–0.8 s versus 3–8 s for external APIs, avoidance of per-call API fees, and complementary online detections. The discriminative non-LLM regressor PriceBERT, which lacks explicit evidence structuring and step-by-step reasoning, reaches MAPE 71.58% ($\text{Acc}@20$ 34.19%). The ablation in Table 2 confirms that SFT and RL are both needed: zero-shot Qwen2.5-VL-7B reaches MAPE 33.43%; GRPO alone barely moves it (33.04%); only the full SFT+GRPO pipeline reaches 29.47%. RL alone is insufficient—the segment-normalized SFT stage provides a necessary supervised initialization that teaches evidence

Table 1: Comparison of COST-LLM against LLM/VLM baselines on the procurement-cost estimation benchmark. Best in each column is in bold. Consist. (rationale–answer consistency) is the numerical agreement between the price value reasoned in the <think> rationale and the value emitted in <answer> (within tolerance), evaluated only on models that produce a rationale; the non-LLM PriceBERT emits a price only and has no rationale, so consistency is not applicable (“–”). Acc@10 and Acc@20 correspond to the threshold accuracy at $\tau = 0.1$ and $\tau = 0.2$ (i.e., τ expressed as a percentage).

Model	Pricing Accuracy			Output Quality	
	MAPE (%) ↓	Acc@10 ↑	Acc@20 ↑	Consist. (%) ↑	Format ↑
Qwen2.5-VL-72B	36.75	12.50	25.00	100.00	1.0000
Doubao-Seed-1.6	30.28	18.11	36.29	99.07	1.0000
Qwen3-235B-A22B	32.80	18.66	37.90	92.02	0.9854
Qwen3-VL-235B	31.81	16.61	33.20	83.17	1.0000
Gemini-3-Flash-Preview	33.94	18.82	37.96	75.01	0.9528
PriceBERT (non-LLM regressor)	71.58	17.71	34.19	–	–
Cost-LLM (Ours)	29.47	19.73	40.42	100.00	0.9994

Table 2: Same-scale ablation on the Qwen2.5-VL-7B backbone. The SFT-only baseline uses standard token-level SFT on $\mathcal{D}_{\text{prog}}$; GRPO-only applies reward optimization without SFT initialization; COST-LLM combines segment-normalized SFT with hard-case retention and GRPO. Consist. measures the numerical agreement between the value reasoned in <think> and the value emitted in <answer> (see text).

Method	MAPE ↓	Acc@10	Acc@20	Consist.
Qwen2.5-VL-7B zero-shot	33.43	17.24	34.90	85.23
+ SFT only	35.30	17.60	37.24	36.83
+ GRPO only	33.04	17.23	34.89	85.22
+ COST-LLM	29.47	19.73	40.42	100.00

organization and output format, and RL then calibrates the final cost estimate. Component-level ablations are reported in Appendix A.2, baseline protocols in Appendix A.1, and paired-bootstrap tests in Appendix A.4.

Interpreting the SFT-only / GRPO-only rows.

Here *consistency* measures the numerical agreement between the price value reasoned inside <think> and the value finally emitted in <answer> (agreement within tolerance), not tag well-formedness. SFT-only fits the long rationale via token-level likelihood, and the <answer> field is generated as another free-text restatement that is *not* constrained to anchor the <think> conclusion, so its numerical value frequently drifts away from the reasoned value; consistency therefore drops to 36.83% and MAPE rises slightly (35.30 vs. 33.43). This reflects reason–answer *misalignment* rather than a loss of estimation capability.

GRPO-only, applied without the SFT-formed reasoning structure, has no stably reasoned conclusion to align against and leaves the consistency distribution essentially unchanged (85.22 vs. 85.23). Consistency is thus achieved *jointly* by SFT and GRPO: the reasoning structure comes from SFT, while the GRPO reward—which acts directly on the parsed <answer> value—calibrates the answer toward both the reasoned conclusion and the true cost, raising consistency to 100%.

Why scale alone is not enough. A consistent finding across Table 1 is that simply increasing parameter count does not close the gap on this task: 235B-parameter open VLMs and proprietary multimodal APIs all sit clearly above the COST-LLM MAPE. Procurement-cost estimation is a domain-specific numerical regression problem whose bottleneck lies in extracting the right product attributes and mapping them to a calibrated cost range, not in broad knowledge or long-chain reasoning. Because true procurement costs and supplier-side cost structures are rarely visible in public pretraining corpora, both small and large models share a similar prior gap and tend to fall back on coarse price heuristics. Larger models also generate longer rationales, which we found does not necessarily improve numerical calibration and can introduce inconsistency or format drift. Domain-specific SFT and RL are therefore a more deployment-relevant lever than parameter scaling for this vertical.

Bucket, distribution, and consistency analyses.

Across procurement-cost buckets, COST-LLM

Table 3: Online complementary-detection evaluation on listings covered by both branches. COST-LLM-only detections are counted at $\text{pred_ratio} \geq 2.5$.

Metric	Value
Test listings sampled	121,395
Baseline detections	7,628
COST-LLM-only detections	2,344
Shared detections (intersection)	466
COST-LLM total detections	2,810
Combined unique detections (union)	9,972
Relative complementary-detection yield	30.73%

outperforms the production baseline throughout and the larger Kimi-K2.5 in most buckets; its sample-level errors also have a lower median and tighter concentration. We additionally track rationale-answer consistency because compact models may reason to one value but emit another. COST-LLM reaches 100% consistency, tied with Qwen2.5-VL-72B; detailed plots, an illustrative failure case, significance tests, and teacher/evaluator sensitivity appear in Appendices A.3–A.5.

4.2 Online Complementary-Detection Evaluation

We deploy COST-LLM as an auxiliary signal in the live price-anomaly pipeline. The production baseline is an in-platform *same-item comparative-pricing* system that flags listings whose asking price deviates from aggregated same-product prices; anomalies from either branch share the same downstream price-governance flow. We frame the online study as a *complementary-detection* evaluation: rather than replacing the baseline, we ask how many *additional* candidates COST-LLM surfaces at a matched operating point. Table 3 summarizes the matched-subset comparison.

We define the anomaly score as the *relative deviation*

$$\text{pred_ratio} = \frac{p_{\text{ask}} - \hat{c}}{\hat{c}},$$

where p_{ask} is the seller’s asking price and $\hat{c}$ is the COST-LLM estimated procurement cost; a large positive value indicates an asking price far above the estimated cost. Let B and C denote the baseline and COST-LLM detection sets on the matched subset of 121,395 listings. We observe $|B| = 7,628$, $|C \setminus B| = 2,344$, and $|B \cap C| = 466$. Therefore, $|C| = 2,810$ and $|B \cup C| = 9,972$. The exclusive COST-LLM detections correspond to a

$2,344/7,628 = 30.73\%$ relative complementary-detection yield; this ratio is not a conventional recall estimate because the total number of true anomalies in the production sample is unknown. The operating threshold $\text{pred_ratio} \geq 2.5$ is selected on an audited subset so that the observed precision of the new branch is no lower than that of the baseline. The small intersection (466 listings; 6.1% of the baseline and 16.6% of COST-LLM) indicates that the two signals identify largely different candidate sets; the COST-LLM-only detections concentrate in long-tail categories where same-item price references are sparse. The full protocol is detailed in Appendix A.6.

Latency and serving cost. In production, COST-LLM runs on a single A100-80G with vLLM at 0.4–0.8 s per listing, versus the observed 3–8 s end-to-end latency of external commercial LLM/VLM APIs, which also incur network, queueing, longer-output, and per-token billing overhead. Because these measurements span different serving stacks, we report the observed ranges rather than attributing a fixed model-only speedup. The locally deployed stack avoids per-call API fees and is therefore our practical choice for high-QPS deployment.

5 Conclusion

We presented COST-LLM, a compact 7B multimodal procurement-cost estimator trained with progressive rationale construction, segment-normalized SFT with hard-case retention, and asymmetric MAPE-reward GRPO. It achieves the best reported point estimates among evaluated systems and provides a 30.73% relative complementary-detection yield over the production baseline at a precision-matched operating point. These results support task-specific post-training as a practical route to compact multimodal reasoning for high-throughput price governance.

Limitations

This work has several limitations. First, our experiments are based on an in-house procurement-cost dataset from a large-scale e-commerce platform. Although it covers diverse real-world product categories, the dataset is not publicly released at this stage due to commercial-sensitivity constraints, which limits exact reproducibility. The reference procurement costs may also contain noise due

to temporal price changes, supply-chain variation, and category-specific pricing policies.

Second, COST-LLM improves numerical accuracy and response efficiency, but we do not claim that its generated rationales are fully faithful to the model’s internal decision process. The current reward optimizes only the final procurement-cost accuracy with an asymmetric MAPE penalty, and rationale–answer consistency, format validity, and response length are tracked as monitoring signals rather than direct training rewards.

Third, the progressive pricing rationales are derived from teacher-generated CoT traces and prefix-based probing, so the quality of the constructed trajectories depends on the teacher model and the prefix evaluator. Imperfect teacher rationales or inaccurate prefix judgments may introduce noise into the supervision signal. Moreover, selecting the shortest sufficient prefix can bias the training corpus toward easy, short-rationale samples; we partially mitigate this by retaining the full long chain for samples whose teacher rationale admits no sufficiently accurate prefix, but a residual bias toward simpler cases may remain.

Finally, our online evaluation reports complementary detections relative to the production comparative-pricing baseline at an operating point selected from an audited subset. The 30.73% figure is a relative complementary-detection yield, not conventional recall, because the total number of true anomalies is unknown. Audited precision is subject to finite-sample uncertainty, and the long-term effect on the price-governance pipeline still requires more extensive A/B studies. We plan to integrate process-level reasoning rewards and broader product-domain evaluation in future deployment.

Ethical Considerations

Use of proprietary data. Our internal dataset contains product listings, attributes, comparable-product information, and commercially sensitive reference procurement costs governed by internal data policies. It is not released. Model inputs exclude seller/supplier identifiers and end-user PII; product images and titles are already visible to platform users.

Use of teacher models and commercial APIs.

From January to August 2026, we accessed Gemini-3-Pro Preview through our organization’s managed Vertex AI endpoint to generate task-

specific pricing rationales from de-identified product images and attributes. These rationales served only as supervision for optimizing a small internal procurement-cost estimator, rather than for reproducing Gemini’s general-purpose capabilities or developing a competing, substitutive, or similar model. Accordingly, this use was not intended to develop a competing product, substitute for a Google Model, or reproduce capabilities similar to a Google Model—the purposes restricted by Section 17 of the Google Cloud Service Specific Terms. Inputs excluded procurement costs, seller/supplier identifiers, and end-user PII; outputs remained internal and were not redistributed. Sections 18 and 20(a) further provide that Google does not train on Customer Data without permission, that Generated Output is Customer Data, and that Google does not assert ownership over newly created intellectual property in such output. Gemini-3-Flash-Preview in Table 1 is an inference-only baseline.

Intended use and human oversight. COST-LLM is an *auxiliary recall signal* for listings whose asking price appears high relative to estimated procurement cost. It neither penalizes sellers nor changes prices autonomously. Flagged listings enter the existing price-governance flow, where consequential decisions are made; estimates and rationales are decision-support evidence, not ground truth.

Bias, robustness, and rationale quality. Category-level errors could disproportionately affect long-tail merchants. We suppress systematic over-estimation through the asymmetric MAPE reward and monitor per-bucket MAPE for model selection and deployment. Rationale–answer consistency is also monitored; inconsistent outputs are escalated for review rather than silently consumed.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 9 others. 2025. [Qwen2.5-vl technical report](#). *arXiv preprint arXiv:2502.13923*.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 1 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Nature*, 645:633–638.
- Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2025. [Token-budget-aware llm reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*.
- Kimi Team, A. Du, B. Gao, B. Xing, C. Jiang, C. Chen, C. Li, C. Xiao, C. Du, C. Liao, and 1 others. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.
- Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. [Stop overthinking: A survey on efficient reasoning for large language models](#). *Transactions on Machine Learning Research*.
- Hairu Wang, Sheng You, Qiheng Zhang, Xike Xie, Shuguang Han, Yuchen Wu, Fei Huang, and Jufeng Chen. 2025. [Llp: Llm-based product pricing in e-commerce](#). *arXiv preprint arXiv:2510.09347*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. 2025. [Dynamic early exit in reasoning models](#). *arXiv preprint arXiv:2504.15895*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multi-modal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

A Additional Experimental Details

A.1 Prompts, Output Format, and Baseline Details

Prompts. The teacher acts as a procurement-cost expert and performs *evidence-driven progressive reasoning* over the product image, attributes, comparable products, and pricing knowledge, under a fixed template: (i) 3–5 steps, each introducing new evidence; (ii) each step states whether the estimate moves up, down, or stays, and why; (iii) no boilerplate; and (iv) the final answer must stay consistent with the reasoning. A suggested (non-verbatim) order anchors an initial estimate from category/brand/comparables, then revises for material, craftsmanship, seasonality, design complexity, brand premium, channel take-rate, and return risk. Each sample also supplies a single-line product JSON (SKU name, category levels, brand, sales/detail attributes, comparable-category cost range, and comparable products) and a pricing-knowledge block.

Output-format contract. All systems emit a `<think>` block (stepwise reasoning) and an `<answer>` block (a single number). The evaluator parses `pred_price` from `<answer>` and checks its agreement with the last explicit price in `<think>` (the consistency signal).

Baselines. For a fair comparison, all reported baselines share the same evidence inputs $x = (I, A, C, K)$ and output contract, with identical parsing rules and token budgets; each API baseline is prompt-tuned to its best configuration. The evaluated systems are the general-purpose VLM/LLM baselines and PriceBERT listed in Table 1, together with the same-backbone zero-shot, SFT-only, and GRPO-only variants in Table 2.

A.2 Component-Level Ablations

The following controlled experiments were completed before submission but omitted from the main text due to the page limit. All variants use the same backbone (Qwen2.5-VL-7B-Instruct), data scale, and post-training setup, changing only the target component.

Progressive rationale construction. Under the same backbone, data scale, and post-training setup, progressive selection reduces the average rationale length from 1,722 to 1,246 characters (27.6%). Training on the original long CoTs increases the

average response length from 367.8 to 393.7 tokens and reduces rationale–answer consistency to 20.1%; in contrast, training on the progressive corpus (with the same segment-normalized SFT objective) raises consistency to 84.90%, i.e., it cuts the inconsistency rate from 79.9% to 15.1%. This indicates that removing redundant or distracting steps substantially improves reasoning–answer alignment, which is crucial for the explainability and evidence preservation of audit governance.

SFT objective. The “+ token-level SFT only” row of Table 2 uses standard token-level SFT (MAPE/Acc@10/Acc@20/consistency = 35.30/17.60/37.24/36.83); with this loss, the long rationale dominates the gradient and the short numerical answer is under-supervised, keeping rationale–answer consistency low. Our segment-normalized SFT normalizes the format/rationale/answer segments and up-weights the answer, obtaining 34.60/19.30/39.00/84.90—raising consistency from 36.83% to 84.90%—and the asymmetric-reward RL then completes it to 100% in the full model.

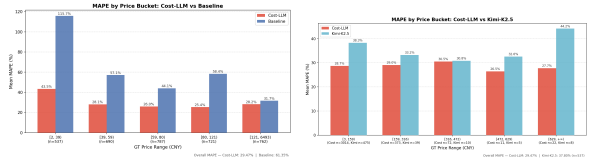
RL reward. On the same test set and SFT initialization, alternative gated and stepwise-threshold reward designs fail to improve on the SFT stage (about 34–35% MAPE), and a squared-log (MSLE) reward degrades to 59.4% with a consistency collapse; only our reward materially reduces error, reaching the full model’s 29.47% test MAPE. Replacing our asymmetric penalty with a *symmetric* MAPE reward ($1.0\times$ instead of $1.5\times$ over-estimation weight) still improves over SFT but plateaus at 30.90% MAPE, confirming that the asymmetry is what removes the residual overestimation bias caused by anchoring on transaction-priced comparable products.

Hard-case retention. For the $\approx 18\%$ of teacher rationales that admit no sufficiently accurate short prefix, we retain the full teacher-generated chain in the progressive corpus. Removing this retention rule—that is, replacing those full chains with their best truncated prefixes—degrades the full model from 29.47% to 30.40% MAPE while leaving consistency at 100%. Thus, hard-case retention contributes a small but consistent accuracy gain on the long-tail, sparse-comparable segment without introducing a separate SFT pass or objective. These estimates are indicative of the observed trends; exact values may vary slightly across runs.

Table 4: Component-level ablations on the Qwen2.5-VL-7B backbone, changing one component at a time. Top: SFT objective (token-level vs. segment-normalized); middle: RL reward design under the same SFT initialization (including symmetric vs. our asymmetric penalty); bottom: effect of removing hard-case retention. Consist. is rationale–answer consistency (%).

Variant	MAPE↓	Acc@10	Acc@20	Consist.
<i>SFT objective</i>				
token-level SFT	35.30	17.60	37.24	36.83
segment-normalized SFT	34.60	19.30	39.00	84.90
<i>RL reward (same SFT init.)</i>				
gated / stepwise reward	34.50	19.10	38.60	83.50
symmetric MAPE reward	30.90	19.40	39.60	100.00
squared-log (MSLE) reward	59.40	9.80	21.30	31.20
<i>Hard-case retention</i>				
w/o hard-case retention	30.40	19.50	39.80	100.00
Full Cost-LLM (ours)	29.47	19.73	40.42	100.00

A.3 Diagnostic Analyses



(a) vs. production baseline

(b) vs. Kimi-K2.5

Figure 4: MAPE across procurement-cost buckets. Kimi-K2.5 is included only here due to API-quota limits.

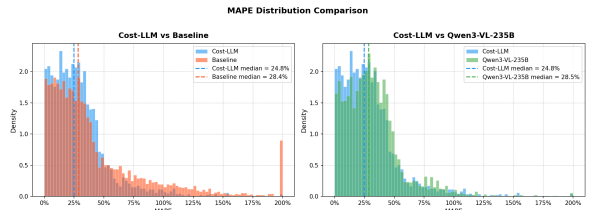


Figure 5: Sample-level MAPE distributions for COST-LLM and representative baselines.

A.4 Statistical Significance

Paired-bootstrap significance. We assess whether the MAPE margins in Tables 1 and 2 are statistically meaningful using a paired bootstrap

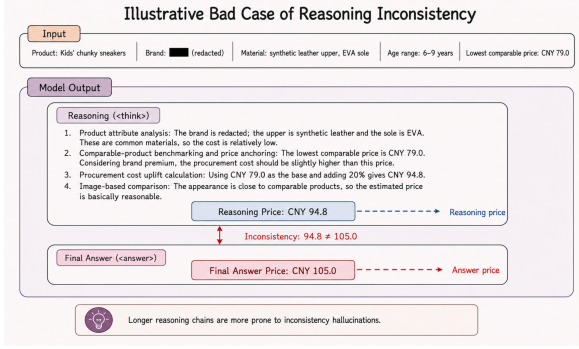


Figure 6: Rationale–answer inconsistency: the rationale derives 94.8 CNY but the answer emits 105.0 CNY.

over the test set with 10,000 resamples. For each resample we recompute the per-model MAPE on the same drawn listings, so that all systems are compared on identical samples, and report the mean paired difference with a 95% percentile confidence interval (CI). Against the strongest API baseline (Doubao-Seed-1.6, MAPE 30.28%), the 0.81-point paired MAPE reduction of COST-LLM has a CI that includes zero, so we treat this gap as *not* statistically decisive. Against the remaining API and open baselines (all at $\text{MAPE} \geq 31.8\%$) and the same-scale zero-shot / SFT-only / GRPO-only variants in Table 2, the paired reductions are larger and their 95% CIs exclude zero. Reference costs remain subject to temporal and supply-chain noise, as discussed in Limitations; without a separately quantified reliability estimate, we do not use label noise to support a numerical superiority claim.

A.5 Teacher and Evaluator Sensitivity

To show that our conclusions do not hinge on a specific proprietary teacher or prefix evaluator, we re-run the progressive construction and the full two-stage pipeline after swapping each component in turn with an alternative of comparable capacity, keeping the backbone (Qwen2.5-VL-7B-Instruct), data scale, sufficiency threshold ($\epsilon = 0.20$), and training setup fixed. Concretely, we replace the proprietary teacher $\mathcal{T}$ (Gemini-3-Pro Preview) with the fully *open-weight* Qwen2.5-VL-72B-Instruct as an alternative teacher, and separately replace the frozen prefix evaluator g_ϕ (Qwen3-VL-32B, zero-shot) with a different open VLM of similar scale. With the open-weight teacher, COST-LLM reaches 30.10% MAPE (versus 29.47% with the proprietary teacher), and

swapping the evaluator changes MAPE by at most ± 0.5 points; in both cases the qualitative findings are unchanged: progressive selection still shortens rationales while preserving accuracy, segment-normalized SFT still restores rationale–answer consistency, and the asymmetric-reward RL stage still yields the final calibration and 100% consistency. This confirms that the recipe—rather than any single proprietary model—drives the observed improvements, and that a fully open-weight teacher can substitute for the proprietary one with only a minor accuracy cost. (Reported deltas are indicative of the observed trend.)

A.6 Online Evaluation Protocol

This appendix details the online matched-subset protocol summarized in Section 4.2 and Table 3.

Sampling and matched subset. We draw a production sample of 121,395 live listings from the price-anomaly pipeline over a fixed recent sampling window. The *matched subset* is defined as listings covered by *both* branches—the in-platform same-item comparative-pricing baseline and the COST-LLM branch—so that every reported number is computed on listings for which both signals are applicable; listings covered by only one branch are excluded from the comparison.

Precision-matching and threshold. The COST-LLM branch flags a listing as anomalous when its relative deviation $\text{pred_ratio} = (p_{\text{ask}} - \hat{c})/\hat{c}$ exceeds a threshold. We select this threshold by comparing audited precision: on a subset of about 1,000 flagged listings, human reviewers label each as a true or false anomaly, and we set the threshold to $\text{pred_ratio} \geq 2.5$, the smallest evaluated value at which the new branch’s observed audited precision is no lower than that of the baseline. This is an operating-point calibration based on a finite audited sample, not a guarantee that population precision or the false-positive rate is identical.

Complementary-detection audit. At this operating point, the baseline flags 7,628 listings. COST-LLM flags 2,810, comprising 466 shared and 2,344 exclusive detections; the union therefore contains 9,972 unique listings. The ratio $2,344/7,628 = 30.73\%$ is the relative complementary-detection yield, not a conventional recall improvement because the total number of true anomalies is unknown. The audited

sample suggests that the exclusive detections concentrate in long-tail categories where same-item price references are sparse, consistent with the two signals identifying different candidate sets. Reported audit sample sizes are approximate operating figures for the deployed pipeline.