The Differential Reasoning Router: Operationalizing Cost-Aware LLM Annotation in E-commerce

Cheng Lyu, Jingyue Zhang, Vinny DeGenova, Mengwei Li, and Yuanli Pei
Wayfair
{clyu1, jzhang2, vdegenova, mli2, ypei}@wayfair.com

Abstract

Large Language Models (LLMs) are increasingly used to annotate structured product data in e-commerce, but early deployment often begins as a cold-start problem: only limited pre-launch labels are available, the value of expensive reasoning is unknown, and human review is needed before the system can be trusted at scale. This challenge is especially common in rule-based annotation workflows, where each item must satisfy multiple business rules and both model errors and ambiguous rule boundaries affect final decisions. We introduce the Differential Reasoning Router (DRR), a cost-aware framework for cold-start LLM annotation that jointly optimizes model selection and human escalation. Rather than treating a reasoning model as a default fallback, DRR estimates separate success probabilities for a direct model and a reasoning model at both the sample and business-rule levels, enabling adaptive routing: easy cases are handled directly, reasoning is reserved for cases where it is expected to improve the decision, and likely double-failure or rule-disagreement cases are escalated to human annotators. The resulting labels provide targeted ground truth for prompt engineering, supervised fine-tuning, calibration, and rule refinement, enabling a gradual shift from human-heavy cold-start annotation toward high-confidence automated routing. In a production e-commerce workflow, DRR reaches accuracy parity with the strongest confidence-based router while achieving more than 60% reasoning-token cost savings.

1 Introduction

In e-commerce, high-quality structured product data are essential for search, recommendation, merchandising, and compliance systems. Maintaining these data often requires rule-based judgments over multimodal evidence, such as whether an image or product description supports a required attribute, whether the displayed product matches the listing, or whether the evidence is too ambiguous for an automated decision. Human reviewers can make these judgments reliably, but catalog review at billion-product scale creates high cost, long latency, and delayed time-to-market for new or updated listings.

Large Language Models (LLMs) and autonomous agents have shown promise in automating parts of this workflow (Mohta et al., 2023; Calderon et al., 2025), and prior work has studied when LLM labels can substitute for or complement human annotation. Existing approaches that select between direct and reasoning models rely heavily on human-labeled datasets. However, new product attributes and evolving business definitions often require immediate catalog-wide deployment before a stable and representative human-annotated set can be constructed. Moreover, human annotations are initially scarce and expensive to obtain. Random sampling may fail to capture challenging or ambiguous cases, and early annotations can quickly become outdated as preliminary rule definitions are refined over time. This creates a cold-start annotation problem: direct prompting is inexpensive but brittle, reasoning can help selectively at higher cost, and the system must allocate traffic across direct inference, additional reasoning, and human review while evidence about the task is still being accumulated. In this context, cold start therefore refers not to the absence of labels, but to a mandatory yet limited pre-launch seed set that is small relative to catalog scale and still evolving with the business rules.

Existing work on model routing improves adaptive computation and test-time compute allocation (Wang et al., 2025; Liang et al., 2025; Ding et al., 2025; Damani et al., 2025), but cold-start annotation introduces two requirements that are not addressed by conventional routing methods alone. First, the router must estimate the *marginal value of reasoning*: whether a more expensive reason-

EMNLP 2026 Industry Track, pages 2815–2829
October 24-29, 2026 ©2026 Association for Computational Linguistics

ing model is likely to change an otherwise direct decision enough to justify its token cost. Scalar confidence and calibration methods can identify uncertain predictions (Kadavath et al., 2022; Ulmer et al., 2024; Khanmohammadi et al., 2025), but they do not directly distinguish examples where reasoning helps from examples where both automated models fail. Second, the router must treat human review as part of the policy. Reasoning models are not reliable universal fallbacks (Shojaee et al., 2025; Zhu et al., 2025), and some failures reflect ambiguous business rules or missing ground truth rather than insufficient model capacity. In these cases, human review is not only a safety mechanism for the current decision, but also a way to acquire labels that reduce future uncertainty.

We propose the **Differential Reasoning Router (DRR)**, a cost-aware framework for cold-start LLM annotation. DRR predicts separate success probabilities for a lightweight Direct model (M_d) and a higher-cost Reasoning model (M_r), both at the sample level and at the level of individual business rules. These differential estimates allow the router to send easy cases to M_d, reserve M_r for cases with positive expected reasoning value, and escalate likely double-failure or rule-disagreement cases to human reviewers. As reviewed labels accumulate, they can be used for prompt iteration, supervised fine-tuning, calibration, and rule refinement, enabling the system to shift gradually from human-heavy cold-start operation toward higher-confidence automated routing.

Our contributions are threefold. First, we formalize cold-start LLM annotation as a joint routing and label-acquisition problem in which the system chooses among direct inference, reasoning, and human review under limited ground truth. Second, we introduce a differentially supervised, budget-constrained routing policy that estimates the marginal value of reasoning while detecting cases that require abstention. Third, we evaluate DRR in a production e-commerce attribute-validation workflow, where it achieves accuracy parity with the strongest confidence-based router while yielding roughly one-fifth higher reasoning-token savings and identifying rule-level disagreement patterns that inform ground-truth collection and rule refinement.

2 Related Work

LLMs as Data Annotators. Mohta et al. (2023) benchmark LLMs as annotators across several tasks and show that downstream models trained on LLM-generated labels consistently underperform those trained on human labels. Complementing this empirical view, Calderon et al. (2025) propose a statistical framework for determining when LLMs can credibly replace human annotators using a limited amount of human labeling. This motivates deployment settings where human labels are scarce but strategically valuable, particularly when a new annotation task is still being calibrated. Domain-specific studies, including financial-data annotation, further show that task expertise, prompt design, and review cost shape whether LLM labels can substitute for human judgments (Aguda et al., 2024). Nevertheless, in domain-specific industrial deployments, hallucinations and reliability concerns often remain a key barrier to adoption. Efforts to mitigate this via calibration and uncertainty estimation, such as probing whether models "know what they know" (Kadavath et al., 2022) or learning confidence signals from model generations and stability under perturbations (Ulmer et al., 2024; Khanmohammadi et al., 2025), are widespread.

Adaptive Computation and Model Routing. Wang et al. (2025) introduce a semantic router for vLLM that predicts a query's reasoning needs and routes it to a cheaper non-reasoning path when appropriate, reducing unnecessary test-time compute. Liang et al. (2025) propose ThinkSwitcher, which dynamically switches a single model between short (fast) and long (slow) chain-of-thought modes based on input complexity. Ding et al. (2025) present BEST-Route, an adaptive routing framework for test-time compute allocation that selects both the model and the best-of-n sampling budget to meet a target quality at minimal estimated cost. Related routing work also studies learning cost-quality trade-offs from data (Damani et al., 2025). While these methods primarily optimize general efficiency, they typically treat the stronger model (or more compute) as a reliable fallback rather than modeling human review as an action for both abstention and label acquisition.

Learning to Defer and Selective Prediction. Learning-to-defer and selective-prediction methods train systems to decide when to defer to an expert rather than make an automated prediction (Madras et al., 2018; Mozannar and Sontag, 2020). These approaches typically study abstention for a fixed predictive model or a fixed human expert under ac-

curacy, fairness, or workload objectives. They do not directly address settings where abstention must be coordinated with model selection between direct and reasoning inference paths, or where deferred examples also serve as ground truth for prompt, rule, and policy refinement during cold-start deployment.

Limits of Reasoning Models. Shojaee et al. (2025) describe an "illusion of thinking", showing that inference-time reasoning effort (*e.g.*, thinking-token compute) can exhibit diminishing or even non-monotonic returns as problem complexity increases. In such cases, accuracy may stagnate or collapse even as models initially spend more compute. Zhu et al. (2025) analyze behavioral divergence under "save-thinking" prompts, showing that reasoning models do not reliably suppress deliberation and can re-engage in explicit or implicit thinking depending on the instance. These findings motivate routing policies that estimate the marginal value of reasoning over direct inference instead of assuming that harder examples should automatically receive more reasoning compute.

3 Methodology

We formulate rule-based product annotation as cost-constrained multimodal routing. Each query $q = (x_{\text{img}}, x_{\text{txt}})$ contains a product image and metadata, and must be evaluated against business rules $\mathcal{R} = \{r_1, \ldots, r_k\}$, where k is the total number of rules. A product is valid only if all rules are satisfied. DRR chooses among a low-cost Direct model (M_d), a higher-cost Reasoning model (M_r), and human review. For each model $m \in \{d, r\}$ and rule r_j, $y_{m,j} = 1$ means M_m agrees with human ground truth on that rule, while $y_m^{\text{sys}} = 1$ means it is correct on the full rule set.

The policy $\pi(q) \in \{M_d, M_r, \text{Human}\}$ must choose among three possible actions: the direct model is sufficient, the reasoning model is worth its extra cost, or human review is needed. Let $C_d(q)$ and $C_r(q)$ denote the model costs, and define the incremental reasoning cost $\Delta C_q = C_r(q) - C_d(q) \geq 0$. The ideal policy minimizes expected error subject to an average reasoning budget B_{target} and a human-referral budget H_{target}:

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E}_q[\text{Risk}(\pi(q), q)] \\ \text{s.t.} \quad & \mathbb{E}_q[\Delta C_q \cdot \mathbf{1}\{\pi(q) = M_r\}] \leq B_{\text{target}}, \\ & \mathbb{E}_q[\mathbf{1}\{\pi(q) = \text{Human}\}] \leq H_{\text{target}}. \end{aligned} \tag{1}$$

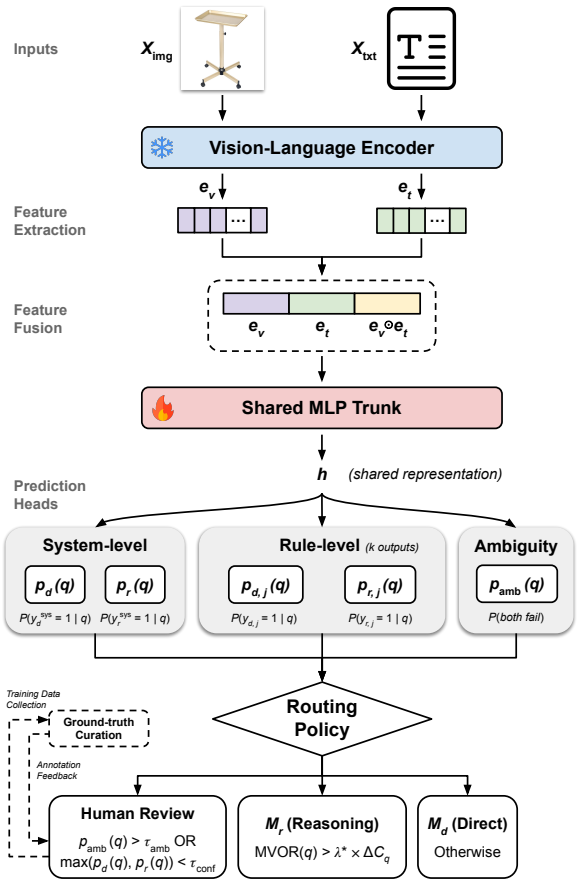

Figure 1: Overall DRR architecture for cost-aware three-way routing with human-feedback support.

We enforce the reasoning budget during training with a nonnegative budget multiplier, and handle human referrals with validation-tuned thresholds.

Figure 1 summarizes the design. Expensive multimodal features are computed once and cached; online routing uses only a small supervised model, which keeps the decision layer compatible with production latency.

3.1 Architecture: Lightweight Differential Predictor

Feature extraction and fusion. We use a pre-trained multimodal encoder to obtain image embeddings e_v and text embeddings e_t. The router does not fine-tune this encoder; it consumes cached features, which keeps routing fast and stable. To capture both unimodal evidence and image-text alignment, we concatenate image, text, and interaction features, $z = [e_v; e_t; e_v \odot e_t]$, then pass z through a shared MLP trunk.

Prediction heads. DRR uses three types of heads. System-level heads predict $\hat{p}_d(q) = P(y_d^{\text{sys}} = 1 \mid q)$ and $\hat{p}_r(q) = P(y_r^{\text{sys}} = 1 \mid q)$, the probabili-

ties that each automated model succeeds on the full rule set. Rule-level heads predict $\hat{p}_{d,j}(q) = P(y_{d,j} = 1 \mid q)$ and $\hat{p}_{r,j}(q) = P(y_{r,j} = 1 \mid q)$, the probabilities that M_d and M_r agree with human ground truth on rule r_j. An ambiguity head predicts $\hat{p}_{\text{amb}}(q)$, the probability that both automated models fail, so the router can escalate cases where more reasoning is unlikely to help.

3.2 Training Objective

DRR is trained to estimate annotator reliability and to spend reasoning budget only where it is expected to improve the decision. For a fixed budget multiplier λ, we combine these goals in an objective minimized over the router parameters θ:

$$\begin{aligned}\mathcal{J}(\theta, \lambda) = \mathcal{L}_{\text{sup}}(\theta) + \mathcal{L}_{\text{route}}(\theta) \\ + \lambda(\bar{C}_{\text{route}} - B_{\text{target}}).\end{aligned} \tag{2}$$

Here $\mathcal{L}_{\text{sup}}$ trains the system-level, rule-level, and ambiguity heads with binary cross-entropy losses; $\mathcal{L}_{\text{route}}$ is the expected error of the relaxed routing policy; and $\bar{C}_{\text{route}}$ is the expected incremental reasoning cost. The multiplier λ acts as the learned price of spending reasoning budget: larger values make routing to M_r harder to justify.

The supervision is differential rather than distillational: DRR learns whether each model matches human ground truth, not whether M_d should imitate M_r. This distinction is important because reasoning can have negative value when M_r fails on a query while M_d gets it correct. The ambiguity target is positive when both system-level correctness labels are zero, i.e., $y_d^{\text{sys}} = 0$ and $y_r^{\text{sys}} = 0$; we use a weighted binary cross-entropy loss for this head because double failures are less common but operationally important.

The routing loss uses the system-level success estimates to compare the two automated actions. For a query that is not sent to human review, the estimated direct and reasoning risks are $1 - \hat{p}_d(q)$ and $1 - \hat{p}_r(q)$, respectively. The central routing quantity is therefore the **Marginal Value of Reasoning** (MVOR), the predicted reduction in risk from using M_r instead of M_d:

$$\text{MVOR}(q) = \hat{p}_r(q) - \hat{p}_d(q).$$

A positive MVOR means reasoning is expected to improve correctness; a negative MVOR means the direct model is expected to be better. During training, the binary choice between M_d and M_r is relaxed with a sigmoid policy so that the routing loss remains differentiable:

$$\pi_r(q) = \sigma\left(\frac{\text{MVOR}(q) - \lambda \Delta C_q}{T}\right),$$

where T is a temperature. The term $\lambda \Delta C_q$ is the opportunity cost of spending reasoning compute on query q.

The relaxed policy then determines both terms in Equation 2: $\mathcal{L}_{\text{route}}$ averages the expected automated error under this mixture of M_d and M_r, while $\bar{C}_{\text{route}}$ averages the corresponding incremental reasoning cost. Thus Equation 2 connects probability learning with budgeted routing: the supervised term learns which model is likely to be correct, while the routing and Lagrangian terms learn how selectively to spend reasoning tokens.

3.3 Optimization

We optimize Equation 2 with alternating primal and dual updates. The model parameters θ are updated by gradient descent, while the multiplier is updated by gradient ascent on the budget violation. With $\lambda = e^{\nu}$, the log-space update is

$$\nu \leftarrow \nu + \eta_\lambda(\bar{C}_{\text{route}} - B_{\text{target}}).$$

If expected reasoning cost exceeds the budget, λ increases and the router becomes more selective; if the model is under budget, λ decreases and more reasoning calls are allowed. Appendix A.5 gives the budget-response and positivity properties of this log-space update, which justify its use as a stable budget price during training.

3.4 Inference Policy

At inference time, DRR uses the learned probabilities and the converged budget multiplier λ^*. It first applies the human gate:

$$\hat{p}_{\text{amb}}(q) > \tau_{\text{amb}} \quad \text{or} \quad \max(\hat{p}_d(q), \hat{p}_r(q)) < \tau_{\text{conf}}.$$

If either condition holds, the query is routed to human review, covering both predicted double failures and broadly uncertain cases.

For the remaining automated cases, the model-selection decision is

$$\pi(q) = \begin{cases} M_r, & \text{if MVOR}(q) > \lambda^* \Delta C_q, \\ M_d, & \text{otherwise.} \end{cases} \tag{3}$$

Appendix A.4 derives this cost-adjusted threshold rule for the automated cases that pass the human gate. The thresholds τ_{amb} and τ_{conf} are tuned

on validation data under a human-referral budget. This gives operators two independent controls: B_{target} governs reasoning spend, while the escalation thresholds govern human review.

3.5 Cold-start Controls and Feedback Loop

The policy is designed to support cold-start deployment and continuous improvement. When task-specific labels are scarce, DRR can initially use conservative human-escalation thresholds to maintain catalog quality while routing uncertain or likely double-failure cases to reviewers. These escalations form an informative label stream rather than a purely random sample: they concentrate on ambiguous examples, cases where M_d and M_r disagree, and instances where both automated paths appear unreliable. As review labels accumulate, DRR can be retrained or recalibrated, prompts or supervised fine-tuning data can be updated, and thresholds can be relaxed to reduce human traffic while retaining an abstention pathway for unreliable cases. Rule-level heads further reveal which business rules underlie disagreement, creating a feedback loop for ground-truth collection and rule refinement.

4 Experiments

4.1 Setup

Task. We evaluate on a production lead-image eligibility task at a major e-commerce platform. Each sample contains a product image and metadata, and the system must decide whether the image can serve as the primary product image. Human ground truth evaluates each sample against k=11 conjunctive business rules spanning objective checks (e.g., no overlaid text, no visible person) and subjective visual judgments (e.g., product is the primary focus, front of product is sufficiently visible). A single rule failure makes the image invalid, so the task is sensitive to both model errors and ambiguous rule interpretation. Appendix A.1 provides the complete rule definitions, and Appendix A.2 shows borderline subjective examples. A sample is *unsolvable* when both M_d and M_r fail on the full rule set, which motivates DRR's human-review action.

Data split and cold-start protocol. The labeled corpus contains 9,358 examples, split into 6,550 training, 1,403 validation, and 1,405 test examples. These labels come from the mandatory pre-launch review process used for launch approval, model/vendor selection, prompt tuning, and operating-point selection; they were not collected as a separate production annotation campaign solely for DRR. This is the sense in which the deployment is cold-start: labels exist, but they are limited relative to catalog scale and still evolving with the business rules. DRR uses the training split to learn task-specific reliability heads, while both DRR and MaxConf use the same validation split to select human-review thresholds. We therefore interpret the comparison as an operational use of required pre-production evidence, not as a label-efficiency claim.

Models. Both automated annotators use Gemini 2.5 Flash Lite with the same structured output prompts; Appendix A.6 provides the full annotation prompts. The Direct model (M_d) runs with thinking disabled, while the Reasoning model (M_r) runs with dynamic thinking enabled, allowing the model to spend additional reasoning tokens when invoked. Note that DRR is not tied to this model version: the router only requires candidate model outputs, realized costs, and ground-truth agreement labels, so newer direct or reasoning models can be substituted by regenerating candidate predictions and recalibrating or retraining the routing layer. We use OpenCLIP ViT-L/14 for feature extraction, with a shared MLP trunk. Embeddings are precomputed and cached.

Baselines. The baselines isolate the design choices in Section 3 using the same model outputs. Direct-only always uses M_d and represents the cheapest automated deployment. Reasoning-only always uses M_r and serves as the reasoning-token cost baseline. Random routes uniformly between M_d and M_r, testing whether learned routing adds value beyond arbitrary model selection. Oracle is a best-of-two automated selector that chooses a correct automated model whenever either M_d or M_r is correct; it shows the limit of automated selection without human review. MaxConf follows confidence-based query-aware LLM selection baselines (Maurya et al., 2025): it routes a sample to human review when the maximum model-reported Direct/Reasoning confidence is below a threshold, and otherwise selects M_r only when its confidence exceeds M_d's. Reasoning-token savings for MaxConf count the incremental reasoning cost only on samples where the policy selects M_r, using the same accounting as DRR. MVOR-only uses the cost-aware M_d/M_r rule without human escalation, isolating the need for abstention.

Method	Acc (%)	Human Review (%)	Reasoning-Token Savings (%)
Direct-only (M_d)	68.0 [65.6, 70.5]	—	100.0
Reasoning-only (M_r)	69.3 [66.8, 71.6]	—	0.0
Random	68.6 [66.5, 70.7]	—	50.0
MaxConf	82.0 [80.0, 84.0]	20.0	55.2
MVOR-only	69.5 [67.1, 71.9]	—	56.4
Oracle	79.1 [77.0, 81.1]	—	86.6
DRR (ours)	**82.8** [80.8, 84.8]	21.7	**66.2**

Table 1: Main results on the product attribute test set (N=1,405). Acc is end-to-end accuracy after routing. Brackets are 95% bootstrap confidence intervals. Reasoning-token savings are measured relative to always using M_r. Thresholded methods are selected on validation set; the table reports realized test-set review rates.

For Pareto sweeps, Ambiguity-only uses only the learned double-failure score for human escalation. Thresholded methods are tuned on the validation set under a $\leq$20% human-referral budget.

Metrics. *End-to-end accuracy* measures whether the final routed decision is correct: automated routes are correct when the selected model matches human ground truth on the full rule set, and human-reviewed samples are counted as correct. *Human review rate* reports the fraction of samples routed to reviewers. *Reasoning token savings* measures reduction in reasoning-token spend relative to always using M_r; it does not include human-review labor. For diagnostic analyses, we distinguish two granularities of double failure: an *unsolvable instance* is a full sample where both M_d and M_r fail the conjunctive decision, while a *rule-level joint failure* is a specific business rule on which both models disagree with human ground truth.

4.2 Main Results

Table 1 addresses the practical question behind the experiment: after allocating limited human review, how much reasoning compute is still needed. The single-model and random rows show that invoking or randomly selecting the reasoning model changes accuracy only modestly. The policies that can route difficult cases to humans reach a higher accuracy range. The discussion below separates these effects into human review, cost-aware reasoning, and ambiguity supervision.

Human review must be a routing action. The methods that always choose between M_d and M_r top out at 79.1% accuracy with Oracle, below the thresholded methods that can route examples to human review. This pattern is expected when some samples are unsolvable by both candidate models: choosing the better automated output is still insufficient for those cases. Human review therefore needs to be modeled as a routing action, rather than added only after automated model selection.

Zone	N Inst.	% Inst.	M_d Acc	M_r Acc	% Unsolv. Inst.
Economy	570	40.6	76.7	78.6	12.8
Reasoning	530	37.7	75.1	79.6	12.5
Ambiguity	305	21.7	39.7	33.8	**50.8**
Population	1405	100	68.0	69.3	20.9

Table 2: Per-zone statistics on the test set. Zones correspond to DRR's three inference actions: direct routing, reasoning routing, and human escalation. Unsolv. Inst. is the percentage of instances in each zone where both automated models fail the full rule set.

Cost-aware routing must be paired with abstention. MVOR-only saves reasoning tokens but remains close to the single-model baselines in accuracy. This separates two effects: the MVOR threshold controls when to spend reasoning tokens, but it does not by itself identify cases that should leave the automated path. After adding the ambiguity gate, DRR reaches the accuracy range of the methods that route examples to humans while achieving more than 60% reasoning-token cost savings compared to MaxConf. Thus cost-aware model selection is most useful when paired with abstention for likely double failures.

Ambiguity supervision adds signal beyond confidence. MaxConf performs well because model-reported confidence is a useful first filter: many difficult examples are also low-confidence examples. However, for cold-start annotation, the review set is also training signal for the next iteration, so the useful cases are not necessarily the least confident ones. The more valuable cases are those that reveal where automation fails, especially examples likely to fail under both M_d and M_r or to expose unstable rule boundaries. DRR uses differential supervision to learn direct-model success, reasoning-model success, and joint automated failure, giving the human gate a task-specific signal beyond confidence alone.

4.3 Routing Zone Analysis

Table 2 checks whether the inference policy creates meaningful regions. The Ambiguity Zone has the highest unsolvable-instance rate, while the Economy and Reasoning Zones retain mostly automat-

Rule	M_d Acc	M_r Acc	% Joint Fail	Type
D1: Primary Focus	68.1	79.4	17.4	Subj.
E1: Front Visible	76.7	78.5	16.7	Subj.
D5: Photorealistic	80.6	81.9	14.5	Subj.
D3: Standard View	79.3	80.1	14.0	Subj.
E3: Correct Set	90.1	84.3	6.3	Subj.
D2: Functionality	95.9	91.0	0.0	Obj.
E2: Standard State	95.2	94.2	0.0	Obj.
E4: Props	96.9	96.7	0.0	Obj.
D4: Dimensions	98.7	98.0	0.0	Obj.
E5: No Text	97.7	99.0	0.0	Obj.
D6: No Person	99.0	99.3	0.0	Obj.

Table 3: Per-rule diagnostics. Joint Fail is the percentage of samples for which both automated models fail the same rule. Subjective rules from Appendix A.2 concentrate these rule-level joint failures, while objective checks are more stable.

able samples. Intuitively, Economy examples tend to have sufficient direct evidence, Reasoning examples require additional cross-checking of the image and metadata, and Ambiguity examples resemble the borderline subjective cases in Appendix A.2. This validates DRR's ability to preserve human capacity while choosing between cheap and expensive model calls.

4.4 Per-Rule Diagnostics

Table 3 explains why the ambiguity region exists: rule-level joint failures concentrate in subjective rules, matching the borderline cases in Appendix A.2. Objective checks are more stable because their criteria are directly verifiable. This validates the rule-level heads in Section 3: they turn system-level routing failures into diagnoses of which business rules need clarification or human oversight. Human referrals therefore serve both as a quality safeguard and as targeted ground truth for future router training and rule refinement.

4.5 Pareto Analysis

Figure 2 treats the human-referral threshold as a deployment control for cold-start annotation. Moving right allocates more reviewer capacity while rules, prompts, and router calibration are still being refined; moving left preserves human capacity but requires more automated decisions. The steep accuracy gains at low review budgets show why selective human escalation is valuable during rollout, and the operating region in Table 1 shows that DRR remains competitive with simple confidence-based escalation while retaining explicit control over reasoning-token spend.

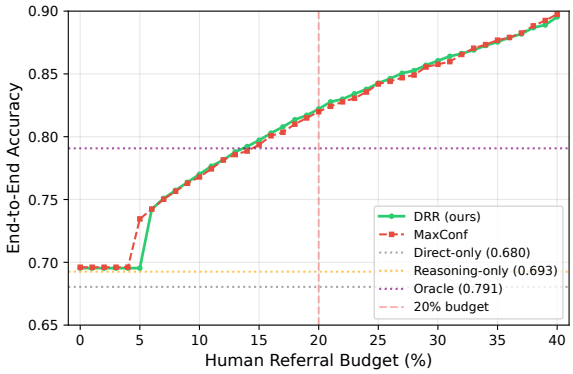

Figure 2: Accuracy-human budget trade-off. The x-axis varies the allowed human-referral budget, and the dashed line marks the operating point used in Table 1. Dotted horizontal lines show automated reference points. The curves show how different escalation policies convert limited human review into end-to-end accuracy during cold-start rollout.

5 Conclusions

DRR shows that routing can do more than select the cheapest successful inference path. By estimating where direct inference is sufficient, where reasoning is likely to help, and where both automated paths are unreliable, the framework turns cold-start LLM annotation into an explicit operating policy. Its routing zones and rule-level diagnostics identify which traffic can be automated, which cases deserve reasoning tokens, and which examples should become human-labeled ground truth. Rather than waiting for a large balanced evaluation set or committing to a single model choice, practitioners can launch with a controlled human-review budget, monitor the accuracy-review tradeoff, and improve prompts, calibration, training data, and rule definitions as the task warms up.

Limitations

This study evaluates DRR on a specific production e-commerce annotation workflow with rule based annotations. Although the framework is intended for rule-based LLM annotations more broadly, its behavior may differ across domains, modalities, rule structures, and operational review processes.

Our experiments also reflect the model landscape at the time of deployment. Although DRR is model-agnostic by design, commercial LLMs evolve rapidly. For each new pair of models, we recommend regenerating candidate predictions and recalibrating or retraining the router to maximize performance gains.

We do not evaluate label efficiency across training-label budgets. The reported DRR result uses the full pre-launch training split, while confidence thresholding requires only validation-set operating-point selection. Thus, our comparison should be read as how to use a mandatory pre-production labeled corpus in deployment, not as evidence that DRR is preferable when all labels must be newly acquired.

Finally, we treat human review as ground truth, matching its role in the production decision process. In practice, human labels can contain disagreement or inconsistency, especially for subjective rules. Such noise may affect both reported accuracy and the supervision used to train ambiguity and rule-level failure signals.

References

Toyin D. Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, and Charese Smiley. 2024. Large language models as financial data annotators: A study on effectiveness and efficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10124–10145, Torino, Italia. ELRA and ICCL.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.

Mehul Damani, Idan Shenfeld, Andi Peng, Andreea Bobu, and Jacob Andreas. 2025. Learning how hard to think: Input-adaptive allocation of lm computation. In *International Conference on Learning Representations*, volume 2025, pages 102783–102802.

Dujian Ding, Ankur Mallick, Shaokun Zhang, Chi Wang, Daniel Madrigal, Mirian Del Carmen Hipolito Garcia, Menglin Xia, Laks V. S. Lakshmanan, Qingyun Wu, and Victor Rühle. 2025. Best-route: adaptive llm routing with test-time optimal compute. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Reza Khanmohammadi, Erfan Miahi, Mehrsa Mardikoraem, Simerjot Kaur, Ivan Brugere, Charese Smiley, Kundan S Thind, and Mohammad M. Ghassemi. 2025. Calibrating LLM confidence by probing perturbed representation stability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10448–10514, Suzhou, China. Association for Computational Linguistics.

Guosheng Liang, Longguang Zhong, Ziyi Yang, and Xiaojun Quan. 2025. ThinkSwitcher: When to think hard, when to think fast. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5185–5201, Suzhou, China. Association for Computational Linguistics.

David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, and Ekaterina Kochmar. 2025. SelectLLM: Query-aware efficient selection algorithm for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20847–20863, Vienna, Austria. Association for Computational Linguistics.

Jay Mohta, Kenan Ak, Yan Xu, and Mingwei Shen. 2023. Are large language models good annotators? In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 38–48. PMLR.

Hussein Mozannar and David Sontag. 2020. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7076–7087. PMLR.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh vahid, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 108018–108059. Curran Associates, Inc.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. 2024. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15440–15459, Bangkok, Thailand. Association for Computational Linguistics.

Chen Wang, Xunzhuo Liu, Yuhan Liu, Yue Zhu, Xiangxi Mo, Junchen Jiang, and Huamin Chen. 2025. When to reason: Semantic router for vllm. *Preprint*, arXiv:2510.08731.

Rongzhi Zhu, Yi Liu, Zequn Sun, Yiwei Wang, and Wei Hu. 2025. When can large reasoning models save thinking? mechanistic analysis of behavioral divergence in reasoning. *Preprint*, arXiv:2505.15276.

A Appendix

A.1 Lead Image Eligibility Rules

The production task studied in this paper is lead-image eligibility: given a product image and listing metadata, decide whether the image is suitable to appear as the primary image shown to customers. The decision is rule-based rather than a single visual-classification judgment. A product image is valid only if it satisfies a conjunction of business rules covering product visibility, presentation state, image cleanliness, product-set consistency, and prohibited visual elements. If any rule fails, the full image-level decision is invalid.

Table 4 lists the 11 rules used for annotation, along with representative examples. The rules span relatively objective visual checks, such as detecting overlaid text, visible people, or dimension markings, and more contextual judgments, such as whether the target product is the primary focus, whether enough of the front is visible, or whether the displayed items match the listed set composition. This mix creates cases where direct inference is sufficient, cases where additional reasoning may help, and cases where human review remains necessary.

A.2 Qualitative Examples of Subjective Rules

Table 5 shows borderline cases for rules with higher subjectivity. In these examples, the visual evidence is present, but the business-rule boundary is underspecified: a product may compete with surrounding room context, be shown at an angle that partially obscures the front, include small edited artifacts, or make set quantity difficult to infer from the image alone. Such cases can produce disagreement between automated annotators and human reviewers, and sometimes between human reviewers themselves.

The examples also clarify why an ambiguity pathway and rule-level diagnostics are useful. Additional reasoning does not always resolve underspecified visual judgments, while rule-level predictions identify which business rule is responsible for a system-level failure. In deployment, reviewed ambiguity cases provide both training signal for future router updates and evidence for refining the underlying rule specification.

A.3 Implementation Details

DRR is implemented as a lightweight supervised routing layer on top of frozen multimodal features. Image and text embeddings are generated offline and cached, so online inference only requires a compact neural router rather than repeated feature extraction. The router shares a small trunk across system-level, rule-level, and ambiguity heads, which keeps the decision layer inexpensive while preserving rule-level diagnostics.

Training uses standard regularized optimization with early stopping on validation performance. The Lagrangian multiplier is updated during training to enforce the reasoning-budget constraint, while human-escalation thresholds are selected on validation data under the target review budget. No additional fine-tuning of the underlying direct or reasoning LLMs is required; all adaptation is concentrated in the routing layer.

A.4 Optimality of Threshold Policy

This section derives the automated model-selection rule used after the human-escalation gate has been applied. It does not claim optimality for the full three-way policy; the human gate is tuned separately under the review budget. Let $\mathcal{A}$ denote the set of queries that pass this gate:

$$\mathcal{A} = \{q : \hat{p}_{\text{amb}}(q) \le \tau_{\text{amb}} \text{ and } \max(\hat{p}_d(q), \hat{p}_r(q)) \ge \tau_{\text{conf}}\}.$$

Conditional on $q \in \mathcal{A}$, the remaining action set is binary, $\{M_d, M_r\}$.

For a query q, define the estimated risks of the direct and reasoning models as

$$R_d(q) = 1 - \hat{p}_d(q), \qquad R_r(q) = 1 - \hat{p}_r(q),$$

and let $\Delta C_q = C_r(q) - C_d(q) \ge 0$ be the incremental cost of invoking reasoning. The Marginal Value of Reasoning is the estimated risk reduction from routing to M_r:

$$\text{MVOR}(q) = R_d(q) - R_r(q) = \hat{p}_r(q) - \hat{p}_d(q).$$

Consider the binary policy $\pi : \mathcal{A} \to \{M_d, M_r\}$ with the same incremental reasoning-budget constraint used in Section 3, restricted to queries that are not sent to human review:

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E}[R_{\pi(q)}(q) \mid q \in \mathcal{A}] \\ \text{s.t.} \quad & \mathbb{E}[\Delta C_q \cdot \mathbf{1}\{\pi(q) = M_r\} \mid q \in \mathcal{A}] \le B_{\text{target}}. \end{aligned} \tag{4}$$

Proposition 1 (Threshold Form of Model Selection)**.** *For any fixed multiplier $\lambda \ge 0$, minimizing the Lagrangian relaxation of Equation 4 over*

Rule	Criterion	Example Type	Example
E1: Front Visible	At least 75% of the product front is visible.	Eligible	
E2: Standard State	The product is shown in its normal, non-functional state.	Eligible	
E3: Correct Set	The image shows the correct quantity or all pieces included in the set.	Eligible	
E4: Props	Props are acceptable only if the target product remains the main focus.	Eligible	
E5: No Text	The image contains no added text, labels, badges, or promotional callouts.	Eligible	
D1: Primary Focus	The target product must be the main visual focus of the image.	Violation	
D2: Functionality	The product should not be shown in use or in an altered functional state.	Violation	
D3: Standard View	The image should show a standard product view, not mainly a side, top, close-up, or detail view.	Violation	
D4: Dimensions	The image should not contain dimensions, arrows, schematics, measurement labels, or similar diagram elements.	Violation	
D5: Photorealistic	The image should be a clean photograph or photorealistic render, without obvious compositing or unrelated inserts.	Violation	
D6: No Person	No person or visible body part should appear in the image.	Violation	

Table 4: Lead image eligibility rules with representative examples. A product image is valid only when all rules are satisfied.

$q \in \mathcal{A}$ decomposes into independent per-query decisions. For each query, the cost-adjusted decision routes to M_r exactly when

$$\mathit{MVOR}(q) > \lambda \Delta C_q, \tag{5}$$

and otherwise routes to M_d.

Proof. For a fixed λ, the Lagrangian relaxation differs from the constrained objective only by the per-query penalty for using M_r:

$$R_{\pi(q)}(q) + \lambda \Delta C_q \cdot \mathbf{1}\{\pi(q) = M_r\}.$$

Because the expectation is linear, the optimal policy can be chosen independently for each query. Routing to M_r is preferred when

$$\begin{aligned} R_r(q) + \lambda \Delta C_q &< R_d(q) \\ R_d(q) - R_r(q) &> \lambda \Delta C_q \\ \mathrm{MVOR}(q) &> \lambda \Delta C_q. \end{aligned}$$

Ties are broken conservatively in favor of M_d, matching the inference rule in Section 3. □

The multiplier λ can be interpreted as a learned shadow price for reasoning budget. Here, "shadow price" simply means the penalty assigned to spending one additional unit of reasoning budget in the routing rule: increasing λ makes reasoning more expensive in the router's decision rule, so fewer queries are sent to M_r. If $\hat{p}_d$ and $\hat{p}_r$ are calibrated,

Table 5: Borderline examples for rules with higher subjectivity. These cases illustrate why human review and rule-level diagnostics remain useful even when reasoning models are available.

Rule	Source of Subjectivity	Example A	Example B
D1: Primary Focus	Requires judging whether the target product dominates the scene. Here, the ***fireplace mantel*** blends into the room and competes with surrounding objects.		
E1: Front Visible	Requires estimating whether enough of the product front is visible. Example A shows the ***fireplace mantel*** at a strong angle, making the visible front area difficult to judge. Example B shows the ***table*** in a staged scene, where surrounding objects make the front view less clear.		
D5: Photorealistic	Requires judging whether small digitally added or composited elements are severe enough to affect image eligibility. Both examples contain minor edited elements, but the overall product presentation remains photorealistic, so these artifacts may be ignored or accepted by human annotators and the model.		
D3: Standard View	Requires judging whether an angled or elevated view still counts as a standard product view. Example A shows a ***three-quarter angled view*** that still represents the product, while Example B uses an elevated room-view angle that may be interpreted differently by annotators.		
E3: Correct Set	Requires judging whether the visible items match the listed quantity or set composition. Example A lists a ***2-piece bedding set***, but appears to show three items. Example B lists ***4 privacy screens/rolls***, but the continuous installation makes the quantity hard to verify.		

this threshold is optimal for true expected risk under the binary model-selection problem; otherwise it is optimal for the estimated risks used by the router. This is why DRR uses supervised probability estimation and validation-based operating-point selection rather than relying on raw model confidence alone.

The same monotonicity appears in the relaxed training policy,

$$\pi_r(q) = \sigma\left(\frac{\mathrm{MVOR}(q) - \lambda \Delta C_q}{T}\right).$$

For fixed temperature $T > 0$, the probability of routing to M_r increases with MVOR and decreases with both incremental cost and the learned budget price λ.

A.5 Dual Update and Constraint Stabilization

DRR uses the Lagrangian term in Equation 2 to discourage routing policies that exceed the reasoning-token budget. Let

$$g(\theta) = \bar{C}_{\text{route}}(\theta) - B_{\text{target}}$$

denote the budget violation of the relaxed routing policy. The primal update treats the current multiplier as fixed and updates the router parameters using

$$\mathcal{L}_{\text{sup}}(\theta) + \mathcal{L}_{\text{route}}(\theta) + \lambda_t g(\theta).$$

The dual variable is then updated in log space:

$$\nu_{t+1} = \nu_t + \eta_\lambda g(\theta_{t+1}), \qquad \lambda_t = e^{\nu_t}. \quad (6)$$

Because the router is a neural model, this section does not claim global convergence to a constrained optimum. Instead, it establishes the exact stabilization properties of the implemented dual update and explains why it is appropriate for enforcing the budget during training.

Proposition 2 (Budget Response of the Log-Space Update)**.** *For any step size $\eta_\lambda > 0$, the update in Equation 6 satisfies*

$$\lambda_{t+1} = \lambda_t \exp(\eta_\lambda g(\theta_{t+1})).$$

Therefore λ_t remains positive for all t; if the expected routing cost exceeds the budget, λ increases; if the router is under budget, λ decreases; and if the budget is exactly met, the dual variable is unchanged.

Proof. Since $\lambda_t = e^{\nu_t}$, exponentiating Equation 6

gives

$$\begin{aligned}\lambda_{t+1} &= e^{\nu_{t+1}} \\ &= e^{\nu_t} \exp(\eta_\lambda g(\theta_{t+1})) \\ &= \lambda_t \exp(\eta_\lambda g(\theta_{t+1}))\,.\end{aligned}$$

The exponential term is always positive, so positivity is preserved. If $g(\theta_{t+1}) > 0$, the exponential factor is greater than one; if $g(\theta_{t+1}) < 0$, it is less than one; and if $g(\theta_{t+1}) = 0$, it equals one. □

Proposition 3 (Mirror-Ascent Interpretation)**.** *For fixed router parameters θ_{t+1}, the multiplicative update in Proposition 2 is the solution to the one-step mirror-ascent problem*

$$\lambda_{t+1} = \arg\max_{u>0} \{\eta_\lambda g(\theta_{t+1})u - D(u\|\lambda_t)\}\,,$$

where $D(u\|\lambda_t) = u\log(u/\lambda_t) - u + \lambda_t$ is the scalar relative-entropy divergence.

Proof. Differentiate the objective with respect to u:

$$\eta_\lambda g(\theta_{t+1}) - \log(u/\lambda_t).$$

Setting this derivative to zero yields $\log(u/\lambda_t) = \eta_\lambda g(\theta_{t+1})$, or

$$u = \lambda_t \exp(\eta_\lambda g(\theta_{t+1}))\,.$$

The objective is strictly concave in u, so the maximizer is unique. □

These propositions give the property needed by DRR: the multiplier is a stable, positive budget price that moves in the correct direction after every batch or epoch. When routing exceeds the target budget, the next primal update penalizes reasoning more strongly; when routing is below budget, the penalty relaxes. This feedback mechanism is sufficient for the role λ plays in the router, namely controlling the threshold $\mathrm{MVOR}(q) > \lambda\Delta C_q$.

Classical primal-dual convergence results require convexity, compactness, and exact or appropriately controlled stochastic updates. Those assumptions do not hold for the non-convex neural router used here, so we do not claim global convergence or optimal constraint satisfaction from the theory alone. In practice, the final operating point is selected on validation data, and the learned multiplier is used only together with validation-tuned human-escalation thresholds.

A.6 Lead Image Evaluation Prompt

Each evaluation uses two prompt components. The system instruction defines the eligibility rules, confidence scoring, and required JSON output. The user message provides the product-specific inputs for each evaluation: the candidate image, the target product, and the catalog text needed to verify quantity, set composition, and visual eligibility. Placeholders such as `[product name]` and `[image]` are replaced with product-specific values before inference.

System Instruction

```
You are an expert AI system performing quality assurance on e-commerce product images.

For each item, you will receive the evaluation rules below, product information, and a candidate lead image.

Instructions:
- Apply all rules exactly as written.
- Do not question or reinterpret the rules or product class.
- Evaluate every rule and provide concise rule-level reasoning.
- If a rule is ambiguous or uncertain, state why and reflect that uncertainty in the confidence score.
- Return only the required JSON structure.

Rule key:
- E = Eligibility requirement. The image must satisfy the rule.
- D = Disqualification condition. The image must not violate the rule.

Rules:

E1_Front_Visible_Enough: The image must show at least 75% of the target product's front, with no major truncation or cropping. The target
      product must be the primary focus. Forward-facing and angled front-facing views are acceptable when the product is visible and
      unobstructed.
Examples: Pass - a coffee table or sofa shown from the front or slight angle, fully in frame. Fail - a product cropped at major edges, less
      than 75% front visible, or visually dominated by foreground objects.

E2_Standard_State: The product must be shown in a standard, non-functional state. In-use styling and minor adjustments are acceptable, such
      as books in a bookcase, decor on a coffee table, an extended luggage handle, or swiveled wheels. Major functional changes that alter
      product form are not acceptable.
Examples: Pass - upright suitcase with handle extended; styled bookcase; coffee table with decor; sofa with pillows. Fail - sofa bed
      unfolded; dresser drawers open; dining table extension pulled apart; storage ottoman with lid removed.

E3_Correct_DSQ_Set: The image must show the correct quantity or all pieces in the set. Use quantity/set information from the product name,
      description, and feature bullets as ground truth, even if it conflicts with the DSQ field. If product text is silent, use DSQ. If
      neither product text nor DSQ specifies quantity, assume one item. Show multiple products only when explicitly specified. Do not treat
      separate items as one unit unless inseparable by design.
Examples: Pass - "Set of 4 Dining Chairs" shows 4 chairs; "10-piece cookware set" shows all 10 pieces; DSQ=4 and product text is silent,
      image shows 4 items. Fail - product text says 3 but image shows 1; set of 2 nightstands shows 1; DSQ=1 but image shows 5 separate sofas.

E4_Props_Appropriate: Props are acceptable only if the target product remains the primary focus.
Examples: Pass - coffee table with mug/book; sofa with pillows that do not obscure it. Fail - chair mostly covered by a prop; dresser mostly
      hidden behind other furniture.

E5_No_Text_In_Image: The image must not contain overlaid or digitally added text, even if it does not obscure the product. Text physically
      printed on the product or visible in decorative background items is allowed when it is not the main focus.
Examples: Pass - wall art text in the background; book titles; branding or printed text that is part of the product design. Fail -
      promotional banner, watermark, URL, measurement callouts, or digitally added "SALE" text.

D1_Not_Primary_Focus: The target product must be the main visual focus. It should occupy a substantial portion of the image, be centrally or
      prominently positioned, and stand out from the surroundings. Visibility alone is not sufficient. If the image is best described as a
      room, setting, or scene with the product present, the product is not the main focus.
Examples: Pass - product centered, large, and prominent, such as a sofa occupying most of the image. Fail - product is small, peripheral, or
      part of a wide scene, such as a light fixture in a porch photo or a chair at the edge of a room scene.

D2_Shows_Functionality: The image must not show the product primarily in a functional or altered-use state.
Examples: Pass - folded towels/sheets; curtains fully closed; styled bookcase. Fail - wardrobe doors/drawers open; refrigerator doors open;
      folding chair folded; shed doors open; chandelier lights on.

D3_Non_Standard_View: The image must not primarily show the side, top, back, close-up, or detail view. Multiple views of one product in a
      single image are not acceptable. Front-facing or slightly angled front views are acceptable when they represent the overall product.
Examples: Pass - sofa shown from front or slight angle, fully visible. Fail - sofa shown from the back only; table shown from top only;
      close-up of armrest/leg; multiple front/back/side views in one image.

D4_Dimensional_Diagram: The image must not contain dimensions, schematics, measurement labels, arrows, or similar diagrams.
Examples: Pass - clean product image with no added dimensions. Fail - product photo with measurement or diagram overlay.

D5_Non_Photorealistic: The image must be a clean, high-quality photograph or photorealistic 3D render. Reject images with obvious pasted
      props, floating accessories, insets, collages, montages, overlays, composited objects, or unrelated floating items. High-quality 3D
      renders are acceptable when indistinguishable from photos.
Examples: Pass - well-lit product photograph; high-quality photorealistic render. Fail - floating product covers/accessories; multiple
      cropped product views in one frame; computer-generated object floating above the product.

D6_Features_Person: The image must not contain a real person.
Examples: Pass - no people present; printed portrait or illustration on wall art or pillow as part of the product design. Fail - person
      sitting on sofa/chair; child sitting on bed.

Final eligibility checklist:
- List all failed rules.
- If any rule fails, the image is not lead eligible.
- If information is contradictory or insufficient, mark lead_image_eligible as "Unsure".

Confidence scoring:
- eligibility_confidence_score must be an integer from 1 to 5.
- 5 = high confidence; all rules and product information are complete, clear, and unambiguous.
- 4 = high confidence; mostly clear with minor ambiguity that does not affect the decision.
- 3 = low confidence; some ambiguity or incomplete information, but a cautious decision is possible.
- 2 = low confidence; significant ambiguity or gaps; decision is a limited-evidence guess.
- 1 = low confidence; information is highly contradictory, mostly missing, or nearly impossible to decide.

eligibility_confidence_reason must be one of:
- High Confidence: reasoning is fully supported and information is present.
- Low Confidence: a major part of the reasoning is unclear or ambiguous.
- Insufficient Information: not enough data to make a strong decision.
- Contradictory Information: image and product text conflict.
- Rule Ambiguous: a rule is open to multiple interpretations.

Output format: Return valid JSON only.

{
  "structured_reasoning": {
    "rules": {
      "E1_Front_Visible_Enough": { "status": "Pass" | "Fail", "reasoning": "..." },
      "E2_Standard_State": { "status": "Pass" | "Fail", "reasoning": "..." },
      "E3_Correct_DSQ_Set": { "status": "Pass" | "Fail" | "Cannot Verify Exception", "reasoning": "..." },
      "E4_Props_Appropriate": { "status": "Pass" | "Fail", "reasoning": "..." },
```

```
      "E5_No_Text_In_Image": { "status": "Pass" | "Fail", "reasoning": "..." },
      "D1_Not_Primary_Focus": { "status": "Pass" | "Fail", "reasoning": "..." },
      "D2_Shows_Functionality": { "status": "Pass" | "Fail" | "Cannot Verify Exception", "reasoning": "..." },
      "D3_Non_Standard_View": { "status": "Pass" | "Fail", "reasoning": "..." },
      "D4_Dimensional_Diagram": { "status": "Pass" | "Fail", "reasoning": "..." },
      "D5_Non_Photorealistic": { "status": "Pass" | "Fail", "reasoning": "..." },
      "D6_Features_Person": { "status": "Pass" | "Fail" | "Cannot Verify Exception", "reasoning": "..." }
    }
  },
  "lead_image_eligible": "True" | "False" | "Unsure",
  "eligibility_confidence_score": int,
  "eligibility_confidence_reason": "High Confidence" | "Low Confidence" | "Insufficient Information" | "Contradictory Information" | "Rule
      Ambiguous",
  "eligibility_correction_suggestion": "..." | null,
  "notes": []
}
```

Product Metadata Template

```
You are evaluating a lead image for the following product:
- **Product Name:** [product name]
## **Product Details**
- **Feature Bullets:** [feature bullets]
- **Product Description:** [description]
- **DSQ (Display Selling Quantity):** [dsq]
- **Set Information:** [set information]

[image]
```