EVIVQA: A DATASET AND BENCHMARK FOR EVIDENCE-GROUNDED EVALUATION OF E-COMMERCE VISUAL QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

E-commerce customer service requires multimodal models to decide not only what to answer, but whether to answer, ask for clarification, abstain, or escalate. Existing benchmarks centered on answer correctness or overall response quality do not fully capture these decisions when evidence is incomplete, conflicting, or beyond available service capabilities. We introduce EviVQA, a Chinese multimodal dataset and benchmark comprising 103,949 instances derived from production e-commerce records. Its retrieval-native inputs combine product images, OCR text, structured attributes, dialogue context, and historical question–answer pairs, while preserving missing information, variant ambiguity, and conflicting evidence. A four-layer annotation scheme connects service scope, evidence sufficiency and information deficits, grounded facts, response constraints, and target actions. These annotations are compiled into decision and evidence certificates, enabling separate evaluation of action selection, factual compliance, and model-judged response quality. On human-annotated, action-balanced examples, we find that higher response-quality scores do not consistently correspond to better action selection, revealing a gap between fluent responses and appropriate customer-service behavior. Combining data curation with certificate-based rewards improves action selection over strong general-purpose and same-backbone baselines, including a 7.6% relative gain in macro-F1 over the strongest fully evaluated general-purpose model. The results further show that improvements in action selection do not necessarily translate into uniform gains in factual compliance or response quality. EviVQA provides a benchmark for studying evidence-grounded customer-service decision making under incomplete information and limited service capabilities.

1 INTRODUCTION

Multimodal customer-service agents answer questions from evidence assembled at inference time, including retrieved product images, OCR text, structured attributes, historical QA, and dialogue context. These sources may omit relevant information, describe different variants, or contradict one another, while some requests, such as checking a replacement shipment, require operational access unavailable from product evidence. As illustrated in Figure 1, reliable assistance therefore requires determining both *what the evidence supports* and *which response action the available evidence and service capabilities justify*.

This decision extends beyond answering or refusing. An agent should **Answer** when evidence supports a useful response, including a qualified one; **Ask** when indispensable user information is missing; **Abstain** when evidence cannot establish the requested fact; and **Escalate** when resolution requires external systems or human authority. Yet response quality alone does not reveal whether the appropriate action was chosen: a fluent reply may promise assistance on a request that requires escalation, while a cautious refusal may withhold supported information. Common alignment pipelines rely on model-generated feedback (Bai et al., 2022) and LLM judges (Zheng et al., 2023), but aggregate quality objectives do not explicitly distinguish these action-level errors.

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

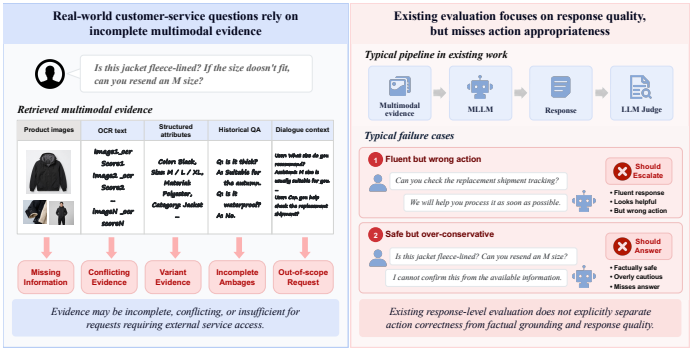


Figure 1: Multimodal evidence is often incomplete or conflicting, while existing evaluation misses action appropriateness.

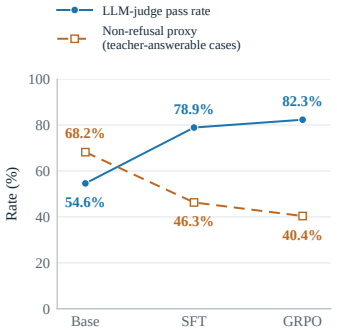


Figure 2: Higher judge pass rates, lower non-refusal rates.

Figure 2 illustrates this mismatch. Across the Base, SFT, and GRPO checkpoints, the LLM-judge pass rate rises from 54.6% to 78.9% and 82.3%, while the non-refusal rate on teacher-answerable cases falls from 68.2% to 46.3% and 40.4%. Thus, higher judge approval can coincide with more frequent refusal on answerable cases, motivating explicit evaluation of action selection alongside factual compliance and response quality.

To address this gap, we introduce **EviVQA**, a retrieval-native Chinese multimodal customer-service dataset and benchmark containing **103,949 instances**, with 101,949 for training and 2,000 for evaluation. Derived from production e-commerce records, it preserves incomplete evidence, variant ambiguity, and service-access constraints rather than assuming every request is answerable. Its four-layer annotations capture **L1** intent and service scope, **L2** evidence sufficiency and deficit attribution, **L3** grounded evidence with verification types and response constraints, and **L4** target actions and risk.

A central distinction in EviVQA is between **answerability and verifiability**. Supported statements may require semantic interpretation, whereas explicit product specifications can often be checked deterministically. We compile eligible annotations into decision and evidence certificates specifying target actions, supported content, explicit negatives, required qualifiers, forbidden claims, and provenance. These certificates enable a **dimension-routed evaluation and reward interface** that separates action matching, factual verification, violation penalties, and response-quality judgments.

We benchmark multimodal models on a balanced held-out test set and conduct a three-seed factorial study of data selection and reward augmentation. Explicit action and factual supervision improves action selection over same-backbone and strong general-purpose baselines, including a **7.6% relative gain in macro-F1** over the strongest fully evaluated general-purpose model. Factorial analysis further shows that reward augmentation contributes the larger average gain in action selection.

Our contributions are:

- We introduce **EviVQA**, a 103,949-instance retrieval-native dataset linking multimodal evidence and service boundaries to **Answer**, **Ask**, **Abstain**, and **Escalate**.
- We develop **four-layer annotations and certificate-based evaluation** that separate action selection, factual compliance, violation detection, and response quality.
- We provide **cross-model benchmarking and three-seed factorial experiments**, showing gains in action selection from certificate-based supervision.

2 Related Work

Reward optimization and model-based feedback. Proxy objectives can improve while intended behavior degrades, as studied in specification gaming and reward-model over-optimization (Amodei et al., 2016; Gao et al., 2023). LLM judges are widely used for evaluating open-ended generations and providing optimization feedback (Zheng et al., 2023; Bai et al., 2022), but their judgments can be biased, uncertain, and stochastic (Fiedler, 2026; Ye et al., 2025; Saha et al., 2026). Similar limita-

Table 1: Comparison of representative QA and customer-service resources. T/I/V: text/image/video; CS: customer service; q.: questions; QA: question–answer pairs; inst.: instances; conv.: conversations; GT: ground truth. Unans. indicates whether answerability annotations are available in at least one task or subset. Grounding denotes explicit links to supporting evidence, not necessarily spatial or temporal localization.

Dataset	Domain	Size	Modality	Unans.	Grounding	Response supervision
Visual Reading and Evidence-Grounded QA						
TextVQA (Singh et al., 2019)	Scene text	45,336 q.	I+T	✗	✗	Answers
DocVQA (Mathew et al., 2021)	Documents	50,000 q.	I+T	✗	✗	Answers
ReceiptQA (Abdalla et al., 2025)	Receipts	171,000 QA	I+T	✓	✗	Answers + answerability
EG-VQA (Huang et al., 2026)	Video QA	11,838 QA	V+T	✗	✓	Answers + temporal/text evidence
E-commerce QA and Multimodal Evaluation						
JD Product QA corpus (Gao et al., 2019)	E-commerce	246,153 samples	T	✗	✗	Answers
FashionVQA (Wang et al., 2022)	Fashion	168M samples	I+T	✗	✗	Attribute answers
Retail VQA case study (Lamm & Keuper, 2024)	Retail	400 prompts	I+T	✗	✗	Feature-value GT (eval.)
EcomMMMU (Ling et al., 2025)	E-commerce	406,190 inst.	I+T	✓	✗	Task targets + answerability
OxyEcomBench (Liu et al., 2026)	E-commerce	∼6.3K inst.	T / I+T	✗	✗	Task-specific reference answers
Shopping Interaction and Customer Service						
MMShopBench (Hao et al., 2026)	Shopping	289 cases	I+T	✗	✓	Verified products + requirements
JourneyBench (Balaji et al., 2026)	Customer service	703 conv.	T	✗	✗	Expected tool / workflow actions (eval.)
EviVQA (ours)	E-commerce CS	103,949 inst.	I+T	✓	✓	Answer / Ask / Abstain / Escalate

tions arise in multimodal settings, where reward models may remain weak on basic visual perception (Li et al., 2025), and coarse response-level preferences can obscure localized factual or behavioral errors (Yu et al., 2024; Wang et al., 2026d). This issue is especially relevant in group-relative RL, where scalar rewards directly determine within-group comparisons and subsequent policy updates (Yu et al., 2025). As a result, improvements in a single aggregate reward may not reveal whether a model is choosing the correct response behavior. We therefore separate action selection from factual compliance and response quality, making action-level errors such as unnecessary refusal or inappropriate answering directly observable.

Verifiability and selective answering. Verifiable rewards are well established in domains such as mathematics and code, where correctness can often be checked automatically, but customer service combines structured product facts with semantic and subjective claims. Recent work demonstrates both the promise and limitations of verifiable supervision (Wang et al., 2026b; Helff et al., 2026). In parallel, unanswerability and abstention have become explicit evaluation objectives (Gurari et al., 2018; Rajpurkar et al., 2018; Zhai et al., 2026; Feng et al., 2024), while recent benchmarks show that strong LLMs may still overcommit under substantial uncertainty (Cocchieri et al., 2026). Customer service further requires distinguishing where the missing information comes from: retrieved evidence may be sufficient, incomplete, absent, or conflicting, while some requests require information from the user or access to external systems. The appropriate action can therefore be to answer, ask for clarification, abstain, or escalate. We further distinguish Ask from Abstain and use counterfactual evidence variants to examine whether model behavior changes with evidential support (Srinivasan et al., 2024; Zhao et al., 2024; Wen et al., 2024).

Multimodal e-commerce QA resources. Existing resources cover complementary parts of this problem. TextVQA and DocVQA focus on visual text and document understanding (Singh et al., 2019; Mathew et al., 2021); FashionVQA, ProductQA, ReceiptQA, and Retail-VQA address product attributes, retail questions, and production-oriented reading (Wang et al., 2022; Gao et al., 2019; Abdalla et al., 2025; Lamm & Keuper, 2024). EcomMMMU extends multimodal e-commerce evaluation to a broader set of tasks (Ling et al., 2025), while shopping-agent benchmarks emphasize interactive workflows, policy following, and tool use (Wang et al., 2026c; Yao et al., 2025). These resources cover visual understanding, product QA, unanswerability, and interactive service at different stages of the customer-service pipeline. As summarized in Table 1, they do not jointly combine retrieval-time multimodal evidence with explicit supervision over **Answer**, **Ask**, **Abstain**, and **Escalate**, source-grounded verification, and counterfactual evidence variants. EviVQA connects these elements within a unified benchmark and compiles eligible annotations into executable evaluation and reward signals.

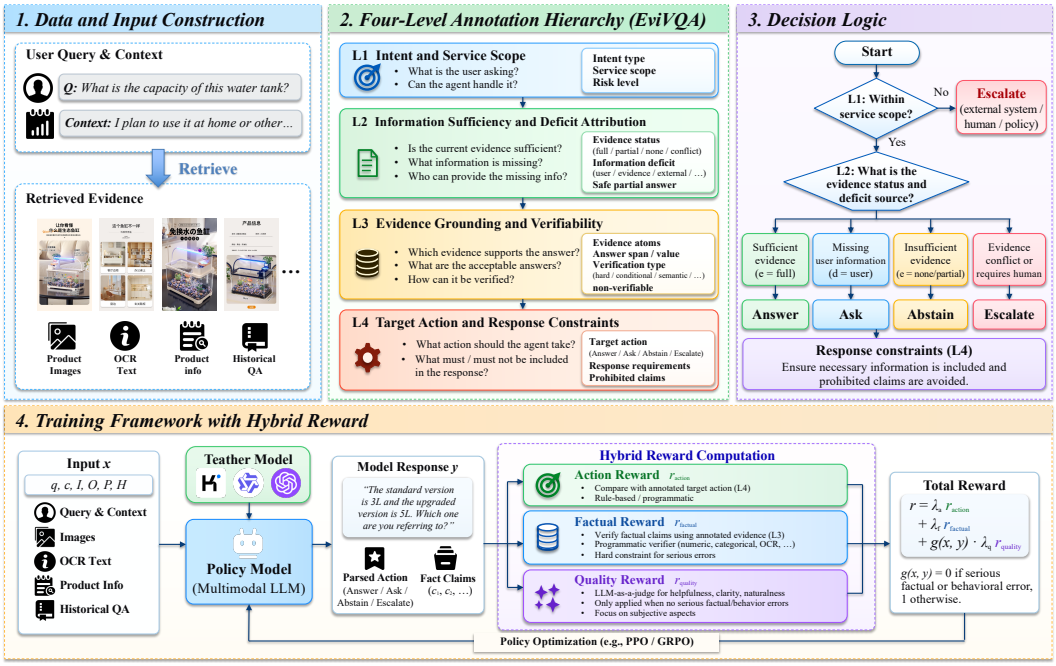


Figure 3: Overview of the EviVQA framework. **Data and Input Construction** assembles retrieval-native multimodal evidence from user queries, dialogue context, product images, OCR text, structured product information, and historical QA; dataset construction and statistics are detailed in Appendix A. **Four-Level Annotation Hierarchy** organizes service scope, evidence sufficiency, grounded evidence and verifiability, and target actions with response constraints; the annotation schema and decision rules are provided in Appendix B. **Decision Logic** maps evidence status and information deficits to **Answer**, **Ask**, **Abstain**, or **Escalate**, as described in Section 3. **Training Framework with Hybrid Reward** converts the resulting annotations into action, factual, and quality signals for policy optimization, with reward design and training detailed in Sections 4 and 4.4.

3 EviVQA: Task Definition and Dataset Construction

EviVQA connects retrieved evidence, service scope, and response actions for customer-service requests. We instantiate the task in EviVQA, derived from production e-commerce records. Figure 3 summarizes the framework.

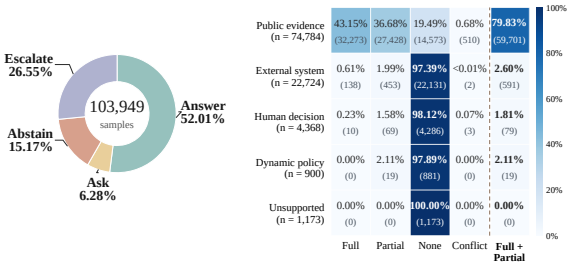
3.1 Retrieval-Native Task and Action Space

Input and evidence. An instance is $x = (q, c, \mathcal{I}, \mathcal{O}, \mathcal{P}, \mathcal{H})$: the user utterance, dialogue context, retrieved images and scores, per-image OCR, structured product information, and historical QA or other product knowledge. The task is *retrieval-native*: supplied evidence may be incomplete, variant-dependent, or conflicting, rather than curated to guarantee answerability. Such settings expose visual-textual grounding and retrieval-reliability challenges in multimodal models (Wu et al., 2026; Ha et al., 2026). Related work also shows that visual evidence can be underused during generation and that counterfactual visual perturbations can reveal reliance on unimodal priors (Favero et al., 2024; Leng et al., 2024). Input characteristics are detailed in Appendix A.2.

Target actions. The target action belongs to $\{\textbf{Answer}, \textbf{Ask}, \textbf{Abstain}, \textbf{Escalate}\}$: **Answer**provides a supported response, including a safe qualified answer; **Ask**requests indispensable user information; **Abstain**withholds unsupported claims; and **Escalate**routes requests requiring external-system access or human authority. This action space distinguishes user-resolvable gaps, insufficient evidence, and service-access constraints instead of treating all non-answer behaviors as refusals.

Table 2: Subsets in EviVQA.

Subset	Annotating Models			Samples
	Qwen	Kimi	GPT	
Bronze	✓	✗	✗	90,797
Silver	✓	✓	✗	11,152
Gold	✓	✓	✓	2000
Total				**103,949**



(a) Target Action Distribution (b) Scope–Answerability Profile

Figure 4: Action and Evidence Distributions.

3.2 Dataset Construction and Annotation Coverage

Source and scale. EviVQA contains 103,949 samples, each pairing a customer-service utterance with available evidence and at least one locally available image. Bronze and Silver constitute the 101,949-sample training split, while Gold forms the 2,000-sample evaluation split. Table 2 summarizes the dataset, with construction details in Appendix A.1.

Annotation coverage. All samples receive canonical annotations from `qwen-vl-max`, while selected subsets receive additional model annotations. Bronze, Silver, and Gold contain one, two, and three or five annotations per sample, respectively. Prompts, model coverage and consensus handling, and cross-model agreement diagnostics are provided in Appendix B.5, Appendix A.3, and Appendix A.5, respectively.

Dataset profile. Figure 4 shows that nearly half of canonical action labels are non-answer actions. Public-evidence requests commonly have full or partial support, whereas external-system and human-decision requests predominantly lack support in the supplied inputs. This contrast motivates separating evidence insufficiency from service-access constraints, consistent with prior work that treats insufficient multimodal context as a detectable condition for abstention (Liu et al., 2024). Distributions by tier, split, scope, and risk appear in Appendix A.4.

3.3 Four-Layer Annotation and Evidence Certificates

Annotation hierarchy. L1 identifies intent and service scope. L2 records evidence sufficiency (`full`, `partial`, `none`, or `conflict`), deficit source, partial-answer safety, and conditional sufficiency. L3 grounds evidence atoms to sources and records applicability conditions and verification types. L4 assigns the target action, risk, and rationale. Partial evidence does not necessarily require **Ask**: a qualified answer may already resolve the question. Definitions, decision rules, and examples appear in Appendix B.

Certificate representation. Beyond action labels, evidence annotations and response constraints specify what a response may claim:

$$C(x) = \big(\mathcal{Y}^+, \mathcal{Y}^-, \mathcal{K}, \mathcal{G}\big), \tag{1}$$

where $\mathcal{Y}^+$ contains supported content, $\mathcal{Y}^-$ explicit negatives and forbidden claims, $\mathcal{K}$ required qualifiers, and $\mathcal{G}$ supporting evidence and provenance. Conditional answering is encoded through $\mathcal{K}$, allowing supported but qualified responses to remain within the **Answer**action. This representation also incorporates hard negatives, which have been shown useful for structured multimodal grounding (Wang et al., 2026a).

Verification and use. `hard` atoms support normalized programmatic checks; `conditional` atoms additionally require matching applicability conditions; `semantic` atoms require evidence-constrained semantic assessment; and `non_verifiable` atoms are excluded from evidence-verification rewards. These verification types enable content-level factual evaluation and structured evidence-grounded training signals, rather than collapsing heterogeneous errors into a single response-level judgment (Wada et al., 2026). Field definitions appear in Appendix B.3.

4 EVIDENCE-GROUNDED EVALUATION AND REWARD DESIGN

The annotations in Section 3 support separate evaluation of action appropriateness, evidence-grounded factual compliance, and response quality. Decision and evidence certificates provide structured signals for the first two dimensions, while model-based judgment assesses the third. These signals are then combined in a dimension-routed reward.

4.1 ANSWERABILITY AND VERIFIABILITY

Answerability concerns whether the available evidence supports an answer, including a safe partial answer, whereas verifiability concerns whether a claim can be checked against that evidence. An evidence-supported answer may still require semantic interpretation, while an explicit product specification can often be verified programmatically. We therefore distinguish evidence support from executable verification.

We denote by $\mathcal{V}$ the instances containing at least one executable `hard` or `conditional` certificate rule, and by $\mathcal{V}_{\text{broad}}$ the broader set that also includes `semantic` verification. Training-time verification operates on executable rules in $\mathcal{V}$, while semantic verification is performed offline. Coverage values are reported in Section 5.

4.2 ACTION RECOGNITION AND CERTIFICATE-BASED VERIFICATION

For a response y, we identify its realized action $\hat{a}$ and extract factual claims for verification against the certificate $C(x)$. This separation follows the broader view that response behavior and answer correctness can be evaluated independently (Gourabathina et al., 2026). When multiple action cues are present, a rule-based classifier assigns the primary action using the priority **Escalate** $>$ **Ask** $>$ **Abstain** $>$ **Answer**. Implementation patterns are provided in Appendix C.2.

The verifier counts certificate-covered unsupported claims (n_u), matches to explicit negatives (n_h), forbidden claims (n_f), missing qualifiers (n_q), and over-refusal I_{or}. The over-refusal indicator is active when the target action is **Answer**, the realized action is **Abstain**, and at least one `hard` evidence atom supports the answer. Samples without an applicable certificate receive a neutral programmatic reward.

4.3 DIMENSION-ROUTED REWARD

Motivated by structured reward design in group-relative optimization (Guo et al., 2026), we combine action matching, factual verification, severe-violation penalties, and response-quality signals:

$$r(x, y) = \lambda_a r_{\text{action}} + \lambda_f r_{\text{factual}} + \lambda_g r_{\text{gate}} + \sum_j \lambda_{q,j} r_{\text{quality},j}. \tag{2}$$

With a matched decision certificate, $r_{\text{action}} = +1$ if $\hat{a} = a^\star$ and -1 otherwise. The factual component deducts weighted penalties for certificate-covered violations:

$$\begin{aligned} r_{\text{factual}} = \text{clip}_{[-1,1]}\big(1 - \alpha_u n_u - \alpha_h n_h - \alpha_f n_f \\ - \alpha_{\text{or}} I_{\text{or}} - \alpha_q n_q\big). \end{aligned} \tag{3}$$

The nonnegative α coefficients weight violation types, while the λ coefficients weight reward components. A factual score of 1 indicates that no certificate-scoped violation was detected, and components without an applicable certificate return 0. Related work has also explored combining evidence sensitivity with verifiable factual correctness (Hou et al., 2026).

The severe-violation component r_{gate} is -1 when $I_{\text{or}} + n_h + n_f > 0$, and 0 otherwise. An LLM judge provides $r_{\text{quality},j}$ for subjective dimensions such as clarity, brevity, and helpfulness. Coefficient settings and implementation details appear in Appendix C.2.

4.4 Training Strategy

We adopt a two-stage training pipeline consisting of supervised fine-tuning followed by group-relative policy optimization. Training hyperparameters, reward coefficients, inference settings, and implementation details are provided in Appendix C.

Stage 1: SFT We fine-tune the base model on 3,687 distilled training instances. Given the user question and available evidence, the model learns to generate the corresponding distilled response. SFT construction and training settings are detailed in Appendix C.1.

Stage 2: GRPO Starting from the shared SFT checkpoint, GRPO refines the policy using relative rewards within groups of sampled responses (Li et al., 2026b). The baseline reward R_0 consists of judge-based signals for format, correctness, grounding, brevity, and helpfulness. The augmented reward R_1 retains these components and adds certificate-derived action and factual rewards, corresponding to r_{action} and r_{factual} in Eq. 2. Reward weights, optimization settings, and the full configuration of R_0 and R_1 are reported in Appendix C.2. The resulting configurations are evaluated in the factorial study described in Section 5.

5 Experimental Evaluation

We evaluate cross-model performance, the contributions of training data and rewards, and the relationship between response quality and action correctness. Experimental details and supplementary analyses are provided in Appendix C.

5.1 Experimental Setup

Data and training. We evaluate on 400 human-annotated held-out items, with 100 targets per action. A 2×2 factorial experiment crosses control and curated data (D_0, D_1) with baseline and augmented rewards (R_0, R_1). Both training sets contain 400 items with matched action counts, while D_1 additionally requires complete certificates. R_1 adds action and factual rewards to R_0, each with weight 0.30; the gate weight is zero. All configurations share the same Qwen3-VL SFT initialization and 50-step LoRA training budget, with three seeds each. Split checks, hyperparameters, and inference settings are detailed in Appendix C.1.

Metrics and reporting. We report action accuracy, macro-F1, refusal rates, factual score, gate violations, and strict JSON compliance. GRPO results average three seed-level evaluations, with macro-F1 computed separately per seed; other models are evaluated once. Under-refusal denotes an **Answer** response to a non-**Answer** target, whereas over-refusal requires an **Answer** target, detected abstention, and hard evidence. Both rates use all collected responses as the denominator. Action recognition, verification rules, and metric aggregation are specified in Appendix C.2. Response quality is scored on a 1–4 scale by `qwen3-vl-plus` using the same rubric v5 as in training; judge and missing-output protocols are described in Appendix C.1.

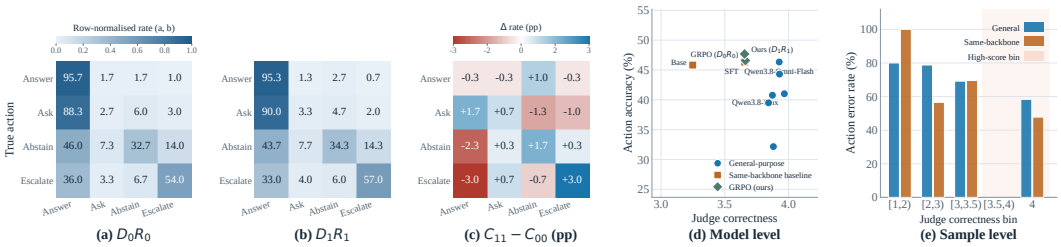
5.2 Main Results

Action-selection performance. Table 3 shows that D_1R_1 achieves the strongest observed action-selection performance among models evaluated on the full 400-item test set, reaching 47.50% accuracy and 42.55% macro-F1. Compared with the strongest fully evaluated general-purpose model, Qwen3.8-Omni-Flash, this corresponds to gains of 1.50 percentage points in accuracy and 3.01 points in macro-F1, or 3.3% and 7.6% relative improvements, respectively. Compared with the same-backbone Base and SFT models, D_1R_1 improves macro-F1 by 2.61 and 1.67 points. The advantage is concentrated in action selection: Omni-Flash has lower under-refusal, while Base achieves the highest factual score and lowest gate-violation rate.

Class-specific gains and remaining errors. Figure 5(a–c) shows that the action gains arise primarily from Escalate and Abstain. Their recall increases from 54.0% to 57.0% and from 32.7% to 34.3%, respectively. By contrast, Ask remains difficult: its recall rises from 2.7% to 3.3%, while

Table 3: Evaluation of multimodal models on EviVQA. GRPO results are arithmetic means across three training seeds; other models are evaluated once. The test set contains 400 items. GPT-5.4 and Claude Sonnet 4.6 results exclude 4 and 1 API-failed items, respectively. Gemini-3.8-Flash is evaluated on a non-stratified 40-item subset and is therefore excluded from the general-purpose average. Rates and macro-F1 are percentages; factual scores use the $[-1, 1]$ scale. Bold indicates the best overall action-selection result, and underlining marks the strongest fully evaluated general-purpose model. The rightmost column reports the mean deviation in action accuracy and macro-F1 from the average of the fully evaluated general-purpose models.

Model	Action Selection		Refusal Behavior		Certificate Checks		Format	Δ Action
	Acc.↑	Macro-F1↑	Over-ref.↓	Under-ref.↓	Factual↑	Gate Viol.↓	JSON↑	vs. GP Avg.↑
General-Purpose Multimodal Models								
GPT-5.4	44.19	37.08	0.00	42.17	0.867	9.34	100.00	+3.95
Claude Sonnet 4.6	32.08	21.63	0.00	59.15	0.864	8.27	45.11	-9.83
Gemini-3.8-Flash	27.50	20.40	0.00	72.50	0.860	10.00	5.00	-12.43
Kimi-K3	39.25	33.09	0.25	50.25	0.877	7.50	96.75	-0.52
Qwen3.8-Max	40.75	32.05	0.25	47.50	0.868	7.50	95.75	-0.29
Qwen3.8-Omni-Flash	46.00	39.54	1.00	37.25	0.874	8.25	99.75	+6.08
Qwen3.7-Plus	40.50	34.07	0.00	47.50	0.877	8.00	95.75	+0.60
Backbone and Supervised Baseline								
Qwen3-VL-4B-Instruct (Base)	45.75	39.94	0.25	40.00	**0.911**	**4.00**	100.00	+6.16
Base + SFT	46.00	40.88	0.50	42.75	0.877	7.75	100.00	+6.75
GRPO-Trained Models								
Base + SFT + GRPO (D_0R_0)	46.25	41.04	0.42	42.58	0.882	7.00	100.00	+6.96
Base + SFT + GRPO (D_1R_1) (Ours)	**47.50**	**42.55**	0.67	41.67	0.880	7.58	100.00	**+8.34**



(a) D_0R_0 (b) D_1R_1 (c) $C_{11} - C_{00}$ (pp) (d) Model level (e) Sample level

Figure 5: Action errors and judge-assessed correctness. (a–b) Row-normalized confusion matrices for D_0R_0 and D_1R_1. (c) $C_{11} - C_{00}$ in percentage points. (d) Model-level action accuracy versus mean judge correctness. (e) Action error rates within judge-correctness bins. D_1R_1 improves Escalate and Abstain recall, while Ask-to-Answer errors remain dominant; high judge scores can still coincide with incorrect actions.

Ask-to-Answer errors increase from 88.3% to 90.0%. The model therefore improves recognition of requests requiring escalation or abstention, while clarification remains the main unresolved decision boundary.

Response quality does not ensure correct actions. Figure 5(d–e) shows that higher judge correctness does not consistently correspond to higher action accuracy. Even among responses receiving the maximum correctness score of 4, action errors occur in more than half of the general-purpose group and approximately half of the same-backbone group. Likewise, 100% JSON compliance among same-backbone models coexists with substantial differences in action accuracy. These results separate response quality and structural compliance from action correctness.

5.3 Ablation Studies

Data and reward contributions. Table 4 reports the four configurations and factorial contrasts. Averaged over the other factor, reward augmentation improves accuracy and macro-F1 by 1.00 and 1.07 percentage points, compared with 0.25 and 0.44 points for data selection. The reward fac-

Table 4: Factorial ablation on the 400-item EviVQA test set. Panel A reports three-seed means. Rates and macro-F1 are percentages; factual scores use the $[-1, 1]$ scale, and response-quality scores use the 1–4 scale. Response quality is evaluated by `qwen3-vl-plus` on all 1,200 responses per configuration. Panel B reports descriptive contrasts in percentage points for rates and macro-F1, and original units for factual and judge scores. Judge contrasts are approximated from the displayed three-decimal means. Bold and underlining mark the best and second-best means in Panel A; JSON compliance is tied.

Configuration / Effect	Action Selection		Refusal Behavior		Certificate Checks		Format	Response Quality			
	Acc.↑	Macro-F1↑	Over-ref.↓	Under-ref.↓	Factual↑	Gate Viol.↓	JSON↑	Correct.↑	Ground.↑	Brevity↑	Helpful.↑
A. Mean Performance across Three Seeds											
D_0R_0 — Baseline	46.25	41.04	**0.42**	42.58	**0.882**	<u>7.00</u>	100.00	<u>3.662</u>	3.942	**3.975**	3.867
D_0R_1 — Reward only	<u>47.08</u>	<u>41.95</u>	0.75	<u>42.50</u>	0.877	7.42	100.00	**3.674**	**3.950**	3.968	**3.878**
D_1R_0 — Data only	46.33	41.31	<u>0.58</u>	42.67	<u>0.881</u>	**6.75**	100.00	3.656	3.943	<u>3.973</u>	3.859
D_1R_1 **— Combined**	**47.50**	**42.55**	0.67	**41.67**	0.880	7.58	100.00	3.654	<u>3.947</u>	3.968	<u>3.867</u>
B. Factorial Effects											
Data effect Δ_D	$+0.25$	$+0.44$	$+0.04$	-0.38	$+0.00067$	-0.04	0.00	-0.0130	-0.0010	-0.0010	-0.0095
Reward effect Δ_R	$+1.00$	$+1.07$	$+0.21$	-0.54	-0.00283	$+0.63$	0.00	$+0.0050$	$+0.0060$	-0.0060	$+0.0095$
Interaction $\Delta_{D\times R}$	$+0.33$	$+0.33$	-0.25	-0.92	$+0.00350$	$+0.42$	0.00	-0.0140	-0.0040	$+0.0020$	-0.0030
Combined − baseline	$+1.25$	$+1.51$	$+0.25$	-0.92	-0.00217	$+0.58$	0.00	-0.0080	$+0.0050$	-0.0070	0.0000

tor therefore contributes the larger average action-selection gain. The interaction is approximately $+0.33$ points for both metrics. Complete contrasts and supplementary mixed-effects analyses are reported in Appendix C.3.

Trade-offs and uncertainty. Relative to D_0R_0, D_1R_1 improves accuracy by 1.25 points and macro-F1 by 1.51 points, while reducing global under-refusal from 42.58% to 41.67%. At the same time, factual score changes from 0.882 to 0.880, gate violations from 7.00% to 7.58%, and over-refusal from 0.42% to 0.67%. The paired hierarchical bootstrap finds no statistically significant factorial contrasts after correction; confidence intervals and procedures are provided in Appendix C.3. We therefore treat these gains as descriptive. Offline recomputation and resource requirements are documented in Appendix C.4.

6 Conclusion

We introduced EviVQA, a Chinese multimodal customer-service dataset and benchmark comprising 103,949 instances derived from production e-commerce records. Designed for a retrieval-native setting, EviVQA captures cases where multimodal evidence may be incomplete, variant-dependent, conflicting, or insufficient for requests requiring external-system access or human intervention. Its four-layer annotation framework connects user intent and service scope, evidence sufficiency and deficit attribution, grounded evidence and verifiability, and target response actions. These annotations distinguish ANSWER, ASK, ABSTAIN, and ESCALATE, and are further compiled into decision and evidence certificates that support separate evaluation of action selection, certificate-scoped factual compliance, and response quality.

Our experiments show that higher response-quality scores do not necessarily correspond to more appropriate action selection, highlighting the limitation of relying on aggregate response-level evaluation alone. Certificate-based supervision improves action-level performance, although these gains are not accompanied by uniform improvements in factual or certificate compliance, suggesting that action appropriateness, evidence grounding, and response quality capture distinct aspects of model behavior. Overall, EviVQA provides a benchmark for studying evidence-grounded decision making under incomplete information and limited service capabilities, and supports the development of customer-service agents that can determine when to answer, ask for clarification, abstain from unsupported claims, or escalate beyond the model's available capabilities. Future work will investigate evidence interventions, richer semantic verification, and more robust clarification behavior.

STATEMENTS

AI USE STATEMENT

Generative AI tools were used to assist with research framing, methodological discussion, translation, drafting, and language editing. They were not used as sources of empirical measurements, experimental results, or ground-truth annotations. All AI-assisted text, numerical claims, citations, figures, and other manuscript artifacts were reviewed and verified by the authors. The authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

ETHICS STATEMENT

EviVQA is derived from real-world e-commerce customer-service records and therefore contains user-generated text and commercially sourced product information. During dataset construction, privacy-sensitive information and personally identifiable information are filtered from releasable data. Product images, metadata, and other commercially sourced content are released only when redistribution is permitted under applicable platform, licensing, and data-governance requirements. Where redistribution is restricted, corresponding content will not be publicly released. The dataset and benchmark are intended solely for research on evidence-grounded multimodal customer-service systems and should not be used to infer private customer attributes, identify individual users, or bypass platform authorization and service boundaries.

REPRODUCIBILITY STATEMENT

The manuscript and appendix specify the task formulation, four-layer annotation framework, target-action decision rules, certificate construction, reward components, evaluation metrics, model configurations, training hyperparameters, data-split construction, and annotation-quality diagnostics used in the reported experiments. The evaluation protocol also documents the human validation procedure for the held-out test set and the handling of model-generated annotations and disagreements. Code, prompts, annotation guidelines, verifier rules, and releasable data artifacts will be made available in anonymized form where permitted by platform, licensing, privacy, and data-governance constraints. For data that cannot be redistributed, we will provide the corresponding processing and evaluation specifications to support reproducibility to the extent permitted.

REFERENCES

Mahmoud Abdalla, Mahmoud SalahEldin Kasem, Mohamed Mahmoud, Bilel Yagoub, Mostafa Farouk Senussi, Abdelrahman Abdallah, Seung Hun Kang, and Hyun Soo Kang. ReceiptQA: A question-answering dataset for receipt understanding. *Mathematics*, 13(11):1760, 2025. doi: 10.3390/math13111760. URL https://doi.org/10.3390/math13111760.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Anthropic. Introducing claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, 2026. System card: https://www.anthropic.com/system-cards.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Sumanth Balaji, Piyush Mishra, Aashraya Sachdeva, and Suraj Agrawal. Beyond IVR: Benchmarking customer support LLM agents for business-adherence. In *Proceedings of the 2026 Conference of the European Chapter of the Association for Computational Linguistics, Industry Track*, 2026.

Alessio Cocchieri, Luca Ragazzi, Giuseppe Tagliavini, and Gianluca Moro. Llms (almost) never abstain under medical uncertainty. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29573–29613. Association for Computational Linguistics, July 2026. doi: 10.18653/v1/2026.acl-long.1365. URL https://aclanthology.org/2026.acl-long.1365/.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Favero_Multi-Modal_Hallucination_Control_by_Visual_Information_Grounding_CVPR_2024_paper.html.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14664–14690. Association for Computational Linguistics, August 2024. doi: 10.18653/v1/2024.acl-long.786. URL https://aclanthology.org/2024.acl-long.786/.

James Fiedler. Bias and uncertainty in llm-as-a-judge estimation, 2026. URL https://arxiv.org/abs/2605.06939.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. Product-aware answer generation in e-commerce question-answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pp. 429–437, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359405. doi: 10.1145/3289600.3290992. URL https://doi.org/10.1145/3289600.3290992.

Google DeepMind. Gemini 3.8 flash model card. https://deepmind.google/models/model-cards/gemini-3-8-flash/, 2026.

Abinitha Gourabathina, Inkit Padhi, Manish Nagireddy, Subhajit Chaudhury, and Prasanna Sattigeri. Answering the wrong question: Reasoning trace inversion for abstention in llms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, July 2026. URL https://aclanthology.org/2026.acl-long.608/.

Peizheng Guo, Jingyao Wang, Wenwen Qiang, Jiahuan Zhou, Changwen Zheng, and Gang Hua. Copo: Causal-oriented policy optimization for hallucinations of mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11051–11063, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Guo_COPO_Causal-Oriented_Policy_Optimization_for_Hallucinations_of_MLLMs_CVPR_2026_paper.html.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018.

Hyeonjeong Ha, Qiusi Zhan, Jeonghwan Kim, Dimitrios Bralios, Saikrishna Sanniboina, Nanyun Peng, Kai-Wei Chang, Daniel Kang, and Heng Ji. Mm-poisonrag: Disrupting multimodal rag with local and global knowledge poisoning attacks. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33804–33826. Association for Computational Linguistics, July 2026. doi: 10.18653/v1/2026.acl-long.1558. URL https://aclanthology.org/2026.acl-long.1558/.

Zeying Hao, Hao Guo, Mengtao Xu, Yimin Hu, Yuheng Song, Zesheng Zhou, Jinsong Lan, and Xiaoyong Zhu. Mmshopbench: A real-log benchmark for multimodal, multi-turn shopping agents, 2026. URL https://arxiv.org/abs/2607.29002.

Lukas Helff, Quentin Delfosse, David Steinmann, Ruben Härle, Hikaru Shindo, Patrick Schramowski, Wolfgang Stammer, Kristian Kersting, and Felix Friedrich. LLMs gaming verifiers: RLVR can lead to reward hacking. *arXiv preprint arXiv:2604.15149*, 2026.

Xuehe Hou, Wenshuo Li, Yali Li, Han Shu, Yuan Wang, Xinghao Chen, and Shengjin Wang. Ves-rft: Rewarding visual evidence sensitivity to mitigate hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4168–4177, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Hou_VES-RFT_Rewarding_Visual_Evidence_Sensitivity_to_Mitigate_Hallucinations_in_Large_CVPR_2026_paper.html.

Linpeng Huang, Weixing Chen, Zexin Chen, Yang Liu, and Liang Lin. EG-VQA: Benchmarking verifiable video question answering with grounded temporal evidence. *arXiv preprint arXiv:2606.24797*, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Bianca Lamm and Janis Keuper. Can visual language models replace OCR-based visual question answering pipelines in production? A case study in retail. *arXiv preprint arXiv:2408.15626*, 2024.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Jinchao Li, Yunxin Li, Chenrui Zhao, Zhenran Xu, Baotian Hu, and Min Zhang. Windowsworld: A process-centric benchmark of autonomous gui agents in professional cross-application environments. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15262–15280. Association for Computational Linguistics, July 2026a. doi: 10.18653/v1/2026.findings-acl.750. URL https://aclanthology.org/2026.findings-acl.750/.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. Vl-rewardbench: A challenging benchmark for vision-language generative reward models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24657–24668, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_VL-RewardBench_A_Challenging_Benchmark_for_Vision-Language_Generative_Reward_Models_CVPR_2025_paper.html.

Ming Li, Pei Chen, Zhenhao Zhang, Tao Yang, Xinyang Zhang, Han Li, Tianyu Cao, Ming Zeng, Zhuofeng Wu, Meng Jiang, Huasheng Li, Lihong Li, and Bing Yin. Mitigating lost in multi-turn conversation via curriculum rl with verifiable accuracy and abstention rewards. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33364–33381. Association for Computational Linguistics, July 2026b. doi: 10.18653/v1/2026.acl-long.1540. URL https://aclanthology.org/2026.acl-long.1540/.

Xinyi Ling, Hanwen Du, Zhihui Zhu, and Xia Ning. EcomMMMU: Strategic utilization of visuals for robust multimodal e-commerce models. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 2025.

Junzhang Liu, Zhecan Wang, Hammad Ayyubi, Haoxuan You, Chris Thomas, Rui Sun, Shih-Fu Chang, and Kai-Wei Chang. Detecting multimodal situations with insufficient context and abstaining from baseless predictions. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8402–8411. ACM, 2024. doi: 10.1145/3664647.3681622. URL https://doi.org/10.1145/3664647.3681622.

Yong Liu, Ximan Liu, Guoqing Yang, Bing Bai, Xiaoqiang Xu, Zhen Chen, Ke Zhang, and Yan Li. OxyEcomBench: Benchmarking multimodal foundation models across e-commerce ecosystems. *arXiv preprint arXiv:2605.13173*, 2026.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021.

OpenAI. Gpt-5.4, 2026. Model release; official model card / announcement URL to be added.

Qwen Team. Qwen3.7-plus, 2026a. Official Qwen blog announcement, June 2026; URL to be added.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork. https://qwen.ai/blog?id=qwen3.8, 2026b.

Qwen Team. Qwen3.8-omni-flash: Omni senses. agentic delivery. https://qwen.ai/blog?id=qwen3.8-omni-flash, 2026c.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 784–789, 2018.

Aadirupa Saha, Aniket Wagde, and Branislav Kveton. Llm-as-a-judge on a budget. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4375–4383. PMLR, 2026. URL https://proceedings.mlr.press/v300/saha26a.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Alok Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12935–12948. Association for Computational Linguistics, August 2024. doi: 10.18653/v1/2024.findings-acl.767. URL https://aclanthology.org/2024.findings-acl.767/.

Yuiga Wada, Kazuki Matsuda, Komei Sugiura, and Graham Neubig. Zina: Multimodal fine-grained hallucination detection and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32528–32538, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wada_ZINA_Multimodal_Fine-grained_Hallucination_Detection_and_Editing_CVPR_2026_paper.html.

Chuhan Wang, Xintong Li, Jennifer Yuntong Zhang, Junda Wu, Chengkai Huang, Lina Yao, Julian McAuley, and Jingbo Shang. Scenealign: Aligning multimodal reasoning to scene graphs in complex visual scenes. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17136–17150. Association for Computational Linguistics, July 2026a. doi: 10.18653/v1/2026.acl-long.779. URL https://aclanthology.org/2026.acl-long.779/.

Jingyi Wang, Lei Zhu, Tengjin Weng, Song-Li Wu, Haochen Tan, Jierun Chen, Chaofan Tao, Haoli Bai, et al. GRPO-VPS: Enhancing group relative policy optimization with verifiable process supervision for effective reasoning. *arXiv preprint arXiv:2604.20659*, 2026b.

Min Wang, Ata Mahjoubfar, and Anupama Joshi. FashionVQA: A domain-specific visual question answering system. *arXiv preprint arXiv:2208.11253*, 2022.

Pei Wang, Yanan Wu, Xiaoshuai Song, Weixun Wang, Gengru Chen, Zhongwen Li, Kezhong Yan, Ken Deng, et al. ShopSimulator: Evaluating and exploring RL-driven LLM agent for shopping assistants. *arXiv preprint arXiv:2601.18225*, 2026c.

Xiaohua Wang, Muzhao Tian, Yuqi Zeng, Zisu Huang, Jiakang Yuan, Bowen Chen, Jingwen Xu, Mingbo Zhou, et al. Reward hacking in the era of large models: Mechanisms, emergent misalignment, challenges. *arXiv preprint arXiv:2604.13602*, 2026d.

Bingbing Wen, Bill Howe, and Lucy Lu Wang. Characterizing llm abstention behavior in science qa with context perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3437–3450. Association for Computational Linguistics, November 2024. doi: 10.18653/v1/2024.findings-emnlp.197. URL https://aclanthology.org/2024.findings-emnlp.197/.

Yixuan Wu, Yang Zhang, Jian Wu, Philip Torr, and Jindong Gu. Postalign: Multimodal grounding as a corrective lens for mllms. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7f52f6b8f107931127eefe15429ee278-Abstract-Conference.html.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b126cc38b8638e07bef37e7b2bb72bf-Abstract-Conference.html.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fdca08d371e4b6c031397909e20043bd-Abstract-Conference.html.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13807–13816, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yu_RLHF-V_Towards_Trustworthy_MLLMs_via_Behavior_Alignment_from_Fine-grained_Correctional_CVPR_2024_paper.html.

Skylar Zhai, Jingcheng Liang, and Dongyeop Kang. Abstain-R1: Calibrated abstention and post-refusal clarification via verifiable RL. *arXiv preprint arXiv:2604.17073*, 2026.

Yulong Zhang, Tianyi Liang, Erfei Cui, Guoqing Wang, Xu Guo, Chenhui Li, and Gongshen Liu. Consensus entropy: Harnessing multi-vlm agreement for self-verifying and self-improving ocr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11643–11653, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_Consensus_Entropy_Harnessing_Multi-VLM_Agreement_for_Self-Verifying_and_Self-Improving_OCR_CVPR_2026_paper.html.

Wenting Zhao, Ge Gao, Claire Cardie, and Alexander M. Rush. I could've asked that: Reformulating unanswerable questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4207–4220. Association for Computational Linguistics, November 2024. doi: 10.18653/v1/2024.emnlp-main.242. URL https://aclanthology.org/2024.emnlp-main.242/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A DATASET: EVIVQA-100K CONSTRUCTION AND DETAILS

This section describes the construction, input characteristics, annotation coverage, label distributions, and annotation-quality diagnostics of EviVQA. Unless otherwise stated, sample-level statistics cover all 103,949 retained samples, and label distributions use the canonical `evics_label` annotations provided by `qwen-vl-max`.

A.1 DATASET CONSTRUCTION

EviVQA is derived from production e-commerce customer-service records. Each sample combines a user utterance with evidence available at retrieval time, including product information, retrieved images, image-associated OCR, and, where available, dialogue history and historical question–answer pairs. The retrieved evidence is not guaranteed to answer the question: missing information, variant ambiguity, service-boundary constraints, and source conflicts remain part of the task.

Table 5 summarizes the dataset. It contains 103,949 distinct samples and 119,161 model-generated annotation records. The latter counts separate model annotations of the same sample and must not be interpreted as the number of task instances. Every retained sample has a non-empty canonical annotation. The dataset contains 101,949 training samples and 2000 evaluation samples. And the evaluation samples come from the Gold tier.

A.2 INPUT CHARACTERISTICS AND IMAGES

Table 6 reports input-length and retrieval statistics. Questions are generally short, with a median of eight characters and a mean of 12.7 characters. Product information is substantially longer, with a median of 472 characters. Dialogue history is non-empty for 44,337 samples (42.65%).

Each sample contains one to five retrieved images, with a median of four, and has at least one locally available image. The retained corpus contains 372,988 image entries, with 372,970 local original-image files. Original images are stored without dataset-level resizing to preserve fine-grained visual and textual details.

All retained image entries have non-empty OCR: 263,464 entries (70.64%) contain at least 50 characters, and 45,971 (12.33%) contain more than 200 characters. These statistics describe the evidence after retrieval cleaning. The upstream retrieval score is unnormalized, with an observed range of 0.394–1.942.

A.3 MODEL-ANNOTATION TIERS AND CONSENSUS

Samples are grouped by annotation coverage: Bronze contains one model annotation per sample, Silver contains two, and Gold contains three or five. The `tier` field records this coverage, while `consensus_status` records agreement. Table 7 summarizes both dimensions. Multi-model annotation covers 13,152 samples (12.65% of the corpus). Within Silver, 7,701 samples (69.05%) have agreement and 3,451 (30.95%) have disagreement. Gold contains 1,208 unanimous samples (60.40%), 684 with a non-unanimous majority (34.20%), and 108 with no majority (5.40%).

Table 5: Overview of the current EviVQA dataset. Indented rows partition the total samples by annotation tier. The training split contains Bronze and Silver samples, while the evaluation split consists of all Gold samples. Image entries include repeated references across samples.

Subset	**Samples**	**Image Entries**
Total samples	**103,949**	**372,988**
Bronze: one annotator	90,797 (87.35%)	325,977
Silver: two annotators	11,152 (10.73%)	39,866
Gold: at least three annotators	2,000 (1.92%)	7,145
Training split (Bronze + Silver)	101,949	365,843
Evaluation split (Gold)	2,000	7,145

Table 6: Input characteristics and retrieval statistics of EviVQA. Text lengths are measured in characters rather than tokens. Statistics are computed over 103,949 samples, except OCR length, which is computed over 372,988 retrieved-image entries.

Metric	Min	P25	Median	Mean	P75	P95	Max
A. Text Lengths (Characters)							
User question	2	6	8	12.7	13	28	478
Dialogue context	0	0	0	262.9	422	1,164	2,785
Product information	17	310	472	501.0	666	962	2,013
OCR per image entry	5	45	75	111.5	130	329	3,385
B. Image and Retrieval Counts per Sample							
Retrieved images	1	3	4	3.6	4	5	5
Locally available images	1	3	4	3.6	4	5	5
Retrieval entries before cleaning	1	5	5	4.95	5	5	5
Retrieval entries after cleaning	1	3	4	3.59	4	5	5
C. Retrieval Score							
Retrieval score	0.394	1.040	1.166	1.17	1.291	1.488	1.942

Table 7: Model-annotation coverage and consensus in EviVQA. Percentages in the last column are calculated within each tier.

Annotation Coverage			Model Agreement	
Tier	**Annotators**	**Samples**	**Consensus Status**	**Count (%)**
Bronze	1	90,797	Single annotation	90,797 (100.00)
Silver	2	11,152	Agreement	7,701 (69.05)
			Disagreement	3,451 (30.95)
Gold	3 or 5	2,000	Unanimous	1,208 (60.40)
			Non-unanimous majority	684 (34.20)
			No majority	108 (5.40)

For two-model disagreements, `consensus_label` is null, whereas `evics_label` retains the Qwen annotation for compatibility. Even for multi-model samples, the canonical annotation is therefore not interchangeable with a consensus annotation.

The six annotators cover 103,949 samples for `qwen-vl-max`, 10,499 for `kimi-k3`, 3,150 for `qwen3.5-omni-plus`, 1,503 for `gpt-5.4-0305-global`, and 30 each for `qvq` and `ds`. These coverage counts overlap. Gold comprises three cohorts: `pilot_5m` ($n = 30$), `eval_omni_3m` ($n = 467$), and `train_gpt_3m` ($n = 1{,}503$).

The Silver-10k annotation round is not identical to the final Silver tier. Silver-10k denotes a fixed cohort of 10,000 samples, whereas the final Silver tier contains samples with exactly two model annotations after all annotation rounds. Samples receiving a third annotation move to Gold.

A.4 Target Actions, Service Scope, and Risk

The four target actions are supported answering (`answer`), requesting necessary user information (`ask`), abstaining from unsupported claims (`abstain`), and transferring requests to an external system or human authority (`escalate`). Table 8 summarizes canonical target-action counts by annotation tier, dataset split, service scope, and risk level. The largest class, `answer`, contains 54,064 samples (52.01%), compared with 6,523 (6.28%) for the smallest class, `ask`, yielding a size ratio of 8.29.

Table 8: Canonical target-action counts by annotation tier, dataset split, service scope, and risk level in EviVQA. Each panel partitions all 103,949 samples. Training consists of Bronze and Silver; evaluation consists of Gold.

Category / Subset	Samples	Answer	Ask	Abstain	Escalate
All samples	**103,949**	**54,064**	**6,523**	**15,764**	**27,598**
A. Annotation Tiers					
Bronze	90,797	49,907	3,830	12,811	24,249
Silver	11,152	3,509	2,291	2,511	2,841
Gold	2,000	648	402	442	508
B. Dataset Splits					
Training	101,949	53,416	6,121	15,322	27,090
Evaluation	2,000	648	402	442	508
C. Service Scope					
Public evidence	74,784	54,052	6,465	14,265	2
External system	22,724	0	2	0	22,722
Human decision	4,368	0	3	0	4,365
Dynamic policy	900	12	1	395	492
Unsupported	1,173	0	52	1,104	17
D. Risk Levels					
Low	61,683	41,815	2,758	10,273	6,837
Medium	32,322	9,754	3,303	3,801	15,464
High	9,944	2,495	462	1,690	5,297

Bronze has a more skewed action distribution, with 54.97% `answer` and 4.22% `ask`. Behavior-balanced sampling increases minority-action coverage in Silver and Gold: the Silver-10k annotation round selected 2,500 samples per canonical action. In the Gold evaluation split, individual action classes account for 20.10–32.40% of samples.

Service scope closely tracks escalation: 22,722 of 22,724 `external_system` samples and 4,365 of 4,368 `human_decision` samples target `escalate`. Risk is annotated separately and spans all four actions. Among the 9,944 high-risk samples, escalation is the most frequent action ($n = 5{,}297$), followed by answering ($n = 2{,}495$), abstention ($n = 1{,}690$), and clarification ($n = 462$).

A.5 Annotation-Quality Diagnostics

We examine annotation consistency through field-level cross-model agreement and deterministic checks of the relationship between service scope, evidence sufficiency, and target action. On the 10,000-sample Silver-10k cohort, `qwen-vl-max` and `kimi-k3` agree on 85.08% of service-scope labels ($\kappa = 0.675$), 68.94% of evidence-sufficiency labels ($\kappa = 0.504$), 66.16% of target-action labels ($\kappa = 0.549$), and 67.12% of risk labels ($\kappa = 0.422$). Here, κ denotes Cohen's kappa. Service scope has the highest agreement, while risk has the lowest chance-corrected agreement. Agreement-based reliability has also been used for training-free OCR verification, where inter-model consensus helps identify sample-level errors (Zhang et al., 2026).

On the 1,503-sample `train_gpt_3m` cohort, target-action agreement is 66.1% between Qwen and Kimi ($\kappa = 0.548$), 67.3% between Qwen and GPT ($\kappa = 0.564$), and 76.0% between Kimi and GPT ($\kappa = 0.665$). Three-model voting yields a majority for 94.34% of samples. Of the 509 Qwen–Kimi action disagreements, the GPT annotation establishes a majority in 424 cases (83.30%). GPT also disagrees with 129 of the 994 labels on which Qwen and Kimi agree (12.98%). The third annotation thus resolves most two-model ties and identifies additional cases of disagreement.

To further assess the reliability of the evaluation labels, we conducted an expert validation of the four-way target-action annotations. Two domain experts manually reviewed the evaluation samples by examining the user query, dialogue context, and available multimodal evidence, and assessed whether each model-generated target action was consistent with the annotation criteria for `answer`, `ask`, `abstain`, and `escalate`. When a model-generated label was judged to be inconsistent with the annotation rules, the reason for disagreement was documented and the label was revised accordingly. Approximately 95% of the model-generated target-action labels were accepted without revision during expert validation. The remaining disagreement cases were manually reviewed and corrected before the evaluation set was used in the final experiments.

Deterministic checks compare each model's target action with the action implied by its own service-scope and evidence-sufficiency annotations under the annotation rules. On the same 1,503-sample cohort, rule-violation rates are 4.13% for Qwen, 4.59% for Kimi, and 8.32% for GPT. The leading mismatch for Qwen is selecting `ask` when the rules imply `answer` (49 cases); for Kimi and GPT, it is selecting `answer` when the rules imply `ask` (68 and 108 cases, respectively). These patterns locate a key source of inconsistency at the boundary between answering and requesting clarification.

B Annotation Schema and Decision Rules

EviVQA organizes annotations into four layers: intent and service scope (L1), evidence sufficiency and deficit attribution (L2), evidence atoms and response constraints (L3), and target action and risk (L4). Together, these layers connect the information required to resolve a request with the evidence available in the input and the appropriate response behavior. This section defines the annotation schema and its decision principles; dataset distributions and consistency diagnostics are reported in Section A.

Table 9 summarizes the annotation fields and their value spaces. Closed fields provide standardized categories for decision-making, while open fields preserve instance-specific evidence and explanations.

B.1 L1: Intent and Service Scope

L1 records the user's intent and assigns a service scope according to the information or authority required to resolve the request. The `scope` field takes one of five values:

- `public_evidence`: requests concerning facts that can, in principle, be established from product descriptions, images, or other supplied reference material.
- `external_system`: requests requiring access to operational records or system functions, such as order status, shipment tracking, or inventory.
- `human_decision`: requests requiring human judgment or authorization, such as discretionary compensation or exception handling.
- `dynamic_policy`: requests governed by changing policies or time-sensitive service rules.
- `unsupported`: requests outside the supported evidence and service capabilities.

Service scope describes where a request can be resolved, rather than whether the current evidence is sufficient. For example, a product-dimension question remains within `public_evidence` even when the retrieved materials omit the requested measurement. Requests for clarification are represented at L4, based on the information gap identified at L2.

B.2 L2: Evidence Sufficiency and Deficit Attribution

L2 assesses the evidence available for the current request. The `sufficiency` field takes one of four values: `full`, when the evidence supports a complete answer; `partial`, when it supports only part of the answer or an answer with explicit conditions; `none`, when it provides insufficient support for a substantive answer; and `conflict`, when relevant sources provide incompatible information.

Table 9: Annotation dimensions and label taxonomy. Closed fields use predefined options; open fields contain evidence-constrained free text or text lists. Structured entries combine multiple subfields. Labels reflect the annotation protocol rather than all observed raw values.

Dimension	Labels or Annotation Content	Value Space
L1. Intent and Service Scope		
User intent	Free-text description of the user's intent	Open
Service scope	Public evidence; external system; human decision; dynamic policy; unsupported	Closed
L2. Evidence Sufficiency and Information Gaps		
Evidence sufficiency	Full; partial; none; conflict	Closed
Deficit source	None; user; public evidence; external system; human; source conflict	Closed
Partial-answer safety	True / false: whether a safe and useful partial answer is available	Closed
Conditional sufficiency	True / false: whether listing variant-specific values and conditions resolves the question	Closed
Deficit note	Free-text description of missing information	Open
L3. Evidence Atoms and Response Constraints		
Evidence source	Image; OCR; product information; QA; context	Closed
Evidence content	Attribute, value, applicability condition, and input-specific image references	Structured
Verification type	Hard; conditional; semantic; non-verifiable	Closed
Acceptable answers	A list of acceptable answer contents	Open
Hard negatives	A list of explicit negative examples	Open
Required qualifiers	A list of conditions or qualifications to preserve	Open
Forbidden claims	A list of claims that must not be asserted	Open
L4. Target Action and Decision Rationale		
Target action	Answer; ask; abstain; escalate	Closed
Risk level	Low; medium; high	Closed
Clarification question	A question requesting necessary user information when the target action is *ask*; otherwise empty	Open
Decision rationale	Free-text explanation referring to specific evidence	Open

The `deficit_source` field identifies the source of the information gap. Its values are `none`, `user`, `public_evidence`, `external_system`, `human`, and `source_conflict`. Here, `none` denotes no identified deficit; the remaining values attribute the gap to missing user information, missing reference evidence, unavailable system records, required human judgment, or conflicting sources, respectively.

Two Boolean fields refine the treatment of partial evidence. `partial_answer_safe`, denoted by p, records whether the available evidence supports a safe and useful partial answer. `conditional_sufficient`, denoted by c_{suff}, records whether explicitly stating the relevant conditions makes the evidence sufficient. For variant-dependent questions, this may involve listing the supported values for each relevant variant. These fields distinguish information that must be requested from the user from information that can be handled through a qualified answer.

B.3 L3: Evidence Atoms and Response Constraints

L3 decomposes the evidence into structured atoms. Each atom records a source, associated image identifiers, an attribute, a value, an applicability condition, and a verification type:

$$z_i = \big(u_i, \mathcal{I}_i, k_i, v_i, \gamma_i, t_i\big), \tag{4}$$

where these components correspond to `source`, `image_ids`, `attribute`, `value`, `condition`, and `verification_type`, respectively. Annotators are instructed to ground factual values in the supplied input and identify their supporting sources. Conditions preserve distinctions such as product variant, operating mode, and applicable usage setting.

Source types and image references. The `source` field specifies the evidence type: `image`, `ocr`, `product_info`, `qa`, or `context`. Image positions are recorded separately in `image_ids`; for example, an atom supported by the second image should use `source=image` with `image_ids=["image_2"]`, rather than `source=image_2`. For image- and OCR-based atoms, the protocol requires non-empty image references pointing to images supplied in the same record.

The schema distinguishes four verification types:

- `hard`: values suitable for deterministic comparison after normalization, such as capacities, dimensions, or numerical limits.
- `conditional`: values that can be checked once an explicit applicability condition is satisfied, such as a capacity associated with a particular SKU.
- `semantic`: evidence-supported statements requiring semantic comparison, such as equivalent descriptions of a care instruction.
- `non_verifiable`: statements that cannot be established from the supplied evidence, such as subjective aesthetic preferences.

Verification types distinguish deterministic comparison, condition-aware checking, semantic assessment, and content that cannot be verified from the supplied evidence. The corresponding verifier implementation is described in Section C.2.

B.4 L4: Target Actions and Decision Rules

L4 assigns a target action from `answer`, `ask`, `abstain`, and `escalate`. These actions correspond to providing an evidence-supported answer, requesting necessary user information, withholding unsupported claims, and transferring the request to an appropriate system or human authority. Conditional answering is represented by `answer` together with response qualifiers, rather than by a fifth action. Risk is recorded separately as `low`, `medium`, or `high`.

We distinguish the model-annotated action a_{ann} from the rule-derived action a_{rule}. Writing service scope, sufficiency, deficit source, and risk as s, e, d, and r, respectively, the decision policy is expressed as

$$a_{\text{rule}} = \pi(s, e, d, p, c_{\text{suff}}, r). \tag{5}$$

The following principles summarize the main decision boundaries, with service requirements considered before evidence-based answering:

1. **Route requests requiring external authority.** Requests with $s \in \{\texttt{external_system}, \texttt{human_decision}\}$ are assigned `escalate`. This boundary is consistent with agent benchmarks showing that cross-application execution is a distinct and difficult capability beyond language-only response generation (Li et al., 2026a).
2. **Request indispensable user information.** When $d = \texttt{user}$ and the missing information is necessary to resolve the request, select `ask`. This applies when a qualified answer cannot already resolve the relevant uncertainty.
3. **Answer within the supported evidence.** For requests within the evidence-answering scope, select `answer` when $e = \texttt{full}$. Partial evidence can also support `answer` when a safe, useful response is available and its limitations or applicability conditions are explicit.
4. **Abstain from unsupported factual claims.** When $e = \texttt{none}$ and the deficit lies in `public_evidence`, select `abstain` rather than infer the missing facts.

5. **Resolve conflicts through the appropriate actor.** For conflicting evidence, request clarification when the user can resolve the ambiguity, escalate when system access or human authority is required, and abstain when the conflict remains unresolved.

Response constraints determine what can be stated within the selected action. An `answer` may require variant-specific qualifiers, an `ask` may include supported background information, and an `escalate` response may explain the required service channel. For rule-consistency diagnostics, a mismatch occurs when $a_{\text{ann}} \neq a_{\text{rule}}$ on an evaluated rule-covered case.

Illustrative example. The case 4 in Figure 8 illustrates the four layers using the request "Please provide the tracking number for the replacement shipment today." L1 assigns `external_system`, because the request requires access to order and shipment records. L2 assigns `none`, with `deficit_source=external_system` and `partial_answer_safe=false`. At L3, the supplied product information and retrieved images provide no evidence for a replacement tracking number. L4 therefore assigns `escalate`, directing the request to a service channel with access to the relevant records. The response can explain this next step without supplying an unsupported tracking number.

B.5 Annotation Prompts and Input Assembly

The annotation interface combines a system prompt defining the L1–L4 task with a record-specific multimodal user message. The user message supplies the question and available evidence. Figures 7 and 6 show English translations of the system prompt and user-message template; the deployed prompts are in Chinese.

System prompt. The system prompt, stored in `evics_label_system.jinja2`, specifies the four-layer annotation task and the instructions applied to each record. Its English translation is presented in Figure 7.

User message. The user-message template combines the current question, dialogue history, product information, and retrieved-image evidence. Its placeholders are populated as follows:

- {`user_question`} contains the current question. If the top-level field is empty, the value is read from the corresponding field in `api_response`.
- {`context`} contains the available dialogue history.
- {`product_info`} contains the product title, category, attributes, specifications, parameters, and related question–answer pairs.
- Each `image_i` entry contains the retrieval score and OCR text associated with one retrieved image, with up to five images per record.

The corresponding images are supplied as multimodal inputs alongside the textual evidence, enabling the annotator to inspect both visual content and OCR transcriptions. The template permits empty OCR fields, although all retained image entries in EviVQA have non-empty OCR after retrieval cleaning.

User_Prompt

Text part (sent as the `text` content block)
```
user_question: {user_question}
context: {context}
product_info: {product_info}
image_1 (recall_score={score_1}) OCR: {ocr_1}
image_2 (recall_score={score_2}) OCR: {ocr_2}
...
image_k (recall_score={score_k}) OCR: {ocr_k}

where:
- `{user_question}` — the user's current question (falls back to the `user_question` field inside `api_response` if the top-level field is empty);
- `{context}` — dialogue context from `api_response` (may be empty);
- `{product_info}` — product title / category / attributes / specifications / parameters / related QA from `api_response`;
- `image_i` blocks — one line per retrieved image, up to `MAX_IMAGES = 5`, each reporting the retrieval `recall_score` and the image's OCR text (OCR may be empty).

Multimodal part (appended after the text block)
The URLs of the same up-to-5 retrieved images (`recall_data_clean[*].url`) are appended as `image_url` content blocks, so the annotator model receives both the raw images and their OCR transcriptions. When the `--no-images` flag is set, only the text part is sent (text-only ablation).

Message structure sent to the annotator
```json
[
{"role": "system", "content": "<evics_label_system.jinja2 rendered>"},
{"role": "user", "content": [
{"type": "text", "text": "<text part above>"},
{"type": "image_url", "image_url": {"url": "<image_1 url>"}},
{"type": "image_url", "image_url": {"url": "<image_2 url>"}},
"... up to 5 images ..."
]}
]
```

Figure 6: English translation of the Chinese user-message template. Placeholders are populated with record-specific text, and the corresponding images are supplied as multimodal inputs.

System _Prompt

You are a senior e-commerce customer-service quality-inspection expert, well versed in product knowledge, customer-service conduct norms, and multimodal evidence verification.
*Your task: for each "product-image VQA customer-service sample," produce a **four-layer decision annotation (L1–L4)**. The annotations will be used simultaneously for: gold-standard labels of the evaluation benchmark, decision certificates for verifiable rewards, and evidence-reliance diagnostics. Strictly follow the specification below and output JSON only.*
Input Field Descriptions
- `user_question`: the user's current question (may be extremely short, contain references, or be a non-interrogative sentence)
- `context`: dialogue context (may be empty; contains historical user/agent messages)
*- `images`: 1–5 product images recalled by a black-box retriever (numbered `image_1`..`image_5`); **being recalled ≠ being relevant to the question***
- OCR text of each image: may be empty or incomplete
- `product_info`: product title, category, attributes, specifications (SKU list), parameters, and related QA (including historical question–answer pairs)
Core Principles
*1. **Evidence sufficiency ≠ image relevance**: a high recall score does not mean the image can answer the question.*
*2. **"Not answering" is often the only correct behavior**: when the images genuinely lack the information, when sources conflict, or when the request exceeds the agent's capability, an honest explanation is more correct than a forced answer.*
*3. **Behavioral boundary**: for requests involving order/logistics/refund processing or human commitments, the correct behavior is to accept and hand off (`escalate`), not to answer.*
*4. **Every annotation must be traceable to concrete evidence** (image/OCR/attributes/QA/context); answer values must come from evidence present in the input — never fabricate answer values from common sense.*
*5. **Common sense may be used to judge whether evidence is sufficient, but must not be used to fill in answers.**
L1 — Intent and Service Boundary*
- `intent`: a one-sentence summary of the user's intent.
- `scope` ∈{`public_evidence`, `external_system`, `human_decision`, `dynamic_policy`, `unsupported`}:
- `public_evidence`: resolvable from public product evidence (images/OCR/attributes/specifications/QA/context);
- `external_system`: requires external-system capabilities such as order, logistics, payment, or refund processing;
- `human_decision`: requires human authority or commitment (reshipment, compensation, customization, discount promises, etc.);
- `dynamic_policy`: real-time dynamic information such as price, promotions, or stock, which cannot be determined from product evidence;
- `unsupported`: outside the scope of service.
*- Note: "needs clarification" is **not** a scope; it is an L4 action.*
L2 — Information Sufficiency and Deficit Attribution
- `sufficiency` ∈{`full`, `partial`, `none`, `conflict`}:
- `full`: the evidence is sufficient to give a definitive answer;
- `partial`: only part can be answered, or qualifications are required;
- `none`: no evidence exists within the permitted scope;
- `conflict`: different evidence sources contradict each other.
- `deficit_source` ∈{`none`, `user`, `public_evidence`, `external_system`, `human`, `source_conflict`}:
- `user`: what is missing is information only the user can provide (specific model, usage scenario, installation environment, etc.);
- `public_evidence`: missing from the public product evidence;
- `source_conflict`: contradiction between sources (e.g., the image says 500ml while the title says 450ml).
*- `partial_answer_safe` ∈{`true`, `false`}: whether a **safe and helpful partial answer** can be given under the current evidence. `partial` ≠ must `ask`.*
*- `conditional_sufficient` ∈{`true`, `false`}: when the answer varies by SKU/variant, whether the question is already resolved **after listing all conditional values** (multiple values do not automatically mean clarification is needed).*
- `deficit_note`: a brief description of the missing information (empty string if there is no deficit).
L3 — Evidence Grounding and Verifiability
- `evidence_atoms`: each evidence fact supporting the answer; array elements are
`{source ∈{image, ocr, product_info, qa, context}, image_ids: ["image_1",...], attribute, value, condition, verification_type}`
- `verification_type` ∈{`hard`, `conditional`, `semantic`, `non_verifiable`}:
- `hard`: numeric/enumerated/string values that can be compared programmatically (e.g., tank capacity = 4.5L; contains the word "flea");
- `conditional`: the answer depends on an SKU/variant condition (e.g., basic version 3L / upgraded version 5L);
- `semantic`: requires semantic understanding to verify (e.g., "hand wash only"; a holistic judgment of safety for children);
- `non_verifiable`: subjective recommendations / style judgments / unsupported inferences.
- `value` must have a source findable in the input (image content also counts; specify `image_ids`).
- `acceptable_answers`: a list of acceptable answers (including synonymous expressions and reasonable approximations, e.g., `["4.5L","4500ml","4.5 liters"]`).
- `hard_negatives`: examples of clearly wrong answers (e.g., `["5L","3L"]`), used as hard negatives.
- `required_qualifiers`: qualifications/conditions that must accompany the answer (e.g., "must distinguish between the basic and upgraded versions").
- `forbidden_claims`: assertions not supported by the evidence (e.g., "all specifications have the same capacity").
L4 — Target Action and Response Constraints
- `target_action` ∈{`answer`, `ask`, `abstain`, `escalate`}.
- Decision rules (applied in order; stop at the first match):
1. `scope` ∈{`external_system`, `human_decision`} → `escalate` (gracefully accept + explain the required information/channel);
2. `deficit_source=user` and the missing condition is necessary → `ask`;
3. `sufficiency=full` → `answer`;
4. `sufficiency=partial` and `partial_answer_safe=true` → `answer` (must include qualifying statements);
5. `sufficiency=none` and `deficit_source=public_evidence` → `abstain`;
6. `sufficiency=conflict` → route according to who can resolve it (user can resolve → `ask`; human required → `escalate`; no one can resolve → `abstain`), and the conflict must be disclosed;
7. otherwise, follow the closest rule and explain.
- `risk` ∈{`high`, `medium`, `low`}: involving safety/health/allergy/compliance/compatibility damage/large commitments → `high`.
- `clarify_question`: if `target_action=ask`, give the ideal clarification question (specific and actionable); otherwise an empty string.
- `decision_reason`: a one-sentence decision rationale that must cite specific evidence (e.g., "image 2's OCR explicitly states 'clamp-on desk 10–70MM'").
Output Format
Output only a single JSON object (no markdown code block, no other text):
```json
{
"L1": {"intent": "...", "scope": "public_evidence"},
"L2": {"sufficiency": "full", "deficit_source": "none", "deficit_note": "", "partial_answer_safe": true, "conditional_sufficient": true},
"L3": {
"evidence_atoms": [{"source": "ocr", "image_ids": ["image_2"], "attribute": "clamp-on desk thickness", "value": "10-70mm", "condition": "", "verification_type": "hard"}],
"acceptable_answers": ["can clamp 10-70mm", "20-30mm is within range, so yes"],
"hard_negatives": ["supports only 60mm"],
"required_qualifiers": ["consult if the desk edge is irregular"],
"forbidden_claims": ["all desk boards can be clamped"]
},
"L4": {"target_action": "answer", "risk": "low", "clarify_question": "", "decision_reason": "image 2's OCR and product_info both state the clamp-on desk thickness is 10-70mm; the user's 2-3cm is within range, so answer directly"}
}
```
Annotation Red Lines
- Speak based on evidence; every key value in an `answer` must have a source findable in `evidence_atoms`.
- When unsure about the behavioral boundary: prefer `ask`; never fabricate an `answer`.
- Anything involving order/logistics/refund processing must always be `escalate`d, even if the product question itself is answerable.
- For harm-related questions (allergy, safety, medical) with insufficient evidence, `abstain` or `ask`, and `risk` must be marked `high`.

Figure 7: English translation of the Chinese system prompt for four-layer annotation.

Case 1



L1 *Scope = public_evidence*

L2 *partial (deficit: user) · risk: medium · safe-partial: True*

L3
- ***[hard]*** *推荐养鱼数量 = 50缸推荐养鱼1-30条2-5cm的鱼儿 (product_info)*
- ***[hard]*** *推荐养鱼数量 = 60缸推荐养鱼1-40条2-5cm的鱼儿 (product_info)*
- ***[hard]*** *鱼缸型号 = 50缸 (image_1, image_1)*
- ***[hard]*** *鱼缸尺寸 = 50缸：49.5x21.5x21.2cm (product_info)*

L4 *ASK*

Case 2



L1 *Scope = public_evidence*

L2 *partial (deficit: public_evidence) · risk: medium · safe-partial: True*

L3
- ***[hard]*** *连接方式 = USB或三节7号电池 (product_info)*
- ***[conditional]*** *供电方式 = USB充电 (product_info)*
- ***[semantic]*** *断电后工作能力 = 在充满电的情况下，即使断电也能继续工作一段时间 (qa)*

L4 *ANSWER*

Case 3



L1 *Scope = public_evidence*

L2 *L2: partial (deficit: public_evidence) · risk: medium · safe-partial: True*

L3
- ***[hard]*** *四步梯总高 = 1.米28 (qa)*
- ***[hard]*** *五步梯总高 = 1.5.米 (qa)*
- ***[hard]*** *六步梯总高 = 1.73.米 (qa)*
- ***[semantic]*** *客服解释使用高度+身高=碰触高度 = 是 (context)*

L4 *ANSWER*

Case 4



L1 *Scope = external_system*

L2 *none (deficit: external_system) · risk: medium · safe-partial: False*

L3
- ***[hard]*** *用户投诉 = 收到的盆变形，包装破损，快递暴力运输 (context)*
- ***[hard]*** *客服请求 = 快递面单发我核实下呢亲 (context)*

L4 *ESCALATE*

Figure 8: Examples of the four-layer annotation schema in EviVQA.

C Experimental Details and Additional Results

This appendix presents the experimental setup, reward implementation, and additional analyses for Table 3. We evaluate models on a human-annotated test set of 400 examples, balanced across four target actions. The controlled experiment follows a 2×2 factorial design over training data and reward configuration, with three seeds per configuration, yielding 12 runs and 4,800 evaluated responses.

C.1 Experimental Setup

Models and checkpoints. All fine-tuned models use Qwen3-VL-4.0-Instruct. The Base model is the original instruction-tuned checkpoint, while SFT uses the merged checkpoint `checkpoint-115-merged`. All GRPO runs start from this SFT checkpoint and train new LoRA adapters with the backbone and vision tower frozen. Each run uses its final one-epoch checkpoint at step 50. We additionally compare seven general-purpose multimodal models.

Training data and evaluation splits. The two training conditions, D_0 and D_1, each contain 400 examples with identical action quotas: 160 **Answer**, 120 **Escalate**, 80 **Abstain**, and 40 **Ask**. Within each action, D_0 samples from the lower half of a certificate-structure richness ranking. D_1 samples from the upper half and additionally requires complete intent, scope, sufficiency, and decision-rationale fields, at least one evidence atom, at least two hard negatives and two forbidden claims, and the relevant action-specific fields. All D_1 examples have matched certificates.

The human-annotated evaluation set contains 100 examples for each target action: **Answer**, **Ask**, **Abstain**, and **Escalate**. The two training sets, evaluation set, and development blocklist have zero cross-split overlap under normalized-question, question-prefix, product-title, and image-content-hash matching. The training conditions share action quotas but differ in category composition, so the data factor represents the complete data-selection policy.

Factorial design and training settings. We cross D_0 and D_1 with reward configurations R_0 and R_1, using seeds 3407, 5501, and 7919. Training uses GRPO for one epoch with LoRA rank 32, $\alpha = 64$, dropout 0.05, and bfloat16 precision. The learning rate is 5×10^{-6}, the warmup ratio is 0.03, and the KL coefficient is $\beta = 0.04$. Each prompt receives eight sampled completions at temperature 0.9, with a maximum completion length of 768 tokens. Each run completes 50 optimizer steps and generates 3,200 training completions. The full settings appear in Table 10.

Inference and judge protocol. All models receive the same questions, evidence text, and ordered images. Reference labels and certificates are used only for scoring. Inference requests one response per example at temperature 0 with a 768-token output budget. Only missing or API-failed responses are retried; invalid JSON and truncated responses are retained.

All 4,800 GRPO responses were successfully collected. GPT-5-4.1 and Claude Sonnet 4.6 returned 396 and 399 responses, respectively. Gemini-3.8-Flash returned a partial set of 40 responses, which is shown for reference and excluded from full-set rankings and the common-intersection analysis.

Response quality is evaluated by `qwen3-vl-plus` using rubric v5, the same judge and rubric used during training. Because multimodal reward models can themselves exhibit substantial visual-evaluation errors (Li et al., 2025), these judge scores are treated as a separate subjective-quality dimension rather than as ground-truth factual correctness. The judge receives the input evidence and candidate response without model, configuration, or seed identifiers. Correctness, grounding, brevity, and helpfulness are scored from 1 to 4. Training maps these scores to $[0, 1]$ through $(s-1)/3$; result tables use the original 1–4 scale. Format is evaluated locally. Table 11 presents the rubric.

C.2 Reward and Verifier Implementation

Target labels and certificates. The target action $a^\star$ is taken from the human annotations of the 400-example evaluation set. Each example is paired with an L3/L3 certificate containing evidence values, hard negatives, required qualifiers, and forbidden claims. All 400 evaluation examples have matched certificates. During training, unmatched prompts receive zero for the certificate-based reward components.

Table 10: Training and inference settings. All twelve GRPO runs share these settings except for the data condition, reward configuration, and seed.

Parameter	Setting
A. GRPO Training	
Framework	ms-swift 3.12.6 (GRPO)
Initialization	SFT `checkpoint-115-merged`
Adaptation	LoRA; backbone and vision tower frozen
LoRA settings	$r = 32$, $\alpha = 64$, dropout 0.05
Precision	bfloat16
Examples / epochs	400 / 1
Optimizer steps	50; final checkpoint used
Batch configuration	8 per device; accumulation 2; 4 GPUs
Generations per prompt	8
Prompts / completions	8 / 64 per optimizer step
Completions per run	3,200
Learning rate / warmup	5×10^{-6} / 0.03
KL coefficient β	0.04
Rollout temperature	0.9
Completion limit	768 tokens
Image budget	602,112 pixels; 1,024 tokens
Reward model	`qwen3-vl-plus`, rubric v5
Judge concurrency	8
Training / data seeds	3407, 5501, 7919
B. Hardware and Software	
GPU hardware	8× NVIDIA RTX PRO 5000, 72 GB each
Allocation	4 GPUs per run; 2 concurrent runs
Peak memory	38.06 GiB per run (smoke runs)
CUDA / driver	12.8 / 580.159.03
Python	3.12.13
PyTorch / Transformers	2.10.0+cu128 / 4.57.1
Accelerate / PEFT	1.13.0 / 0.18.1
C. Evaluation Inference	
Local engine / batch	Transformers / 4
Temperature / top-p	0.0 / 1.0
Seed / responses	0 / 1 per input; no best-of
Output limit	768 new tokens
Retry policy	Missing or API-failed responses only
Retained responses	Successful and invalid-JSON outputs
GRPO outputs	$12 \times 400 = 4{,}800$; all collected
Inference time	$\sim$80 minutes for 12 runs (estimated)

Action recognition and factual checks. A rule-based classifier identifies the realized action $\hat{a}$ from the response's primary clause. Escalation cues take precedence over clarification cues. For hedged primary clauses, the cue search extends to the full response; a hedged response without either cue is classified as **Abstain**. An unhedged primary clause without either cue is classified as **Answer**, and an empty clause as **Abstain**.

The verifier checks `hard` atoms using unit-normalized numeric, range, and substring comparisons. For `conditional` atoms, non-empty conditions are additionally checked through keyword matching. Semantic and non-verifiable atoms are outside the programmatic scoring path. The evaluation certificates contain 1,132 atoms: 790 hard, 12 conditional, 307 semantic, and 23 non-verifiable.

Unsupported-claim deductions apply when a non-abstaining response asserts an attribute but fails its certificate check. Hard negatives and forbidden claims are checked by substring matching. Required

Table 11: The five-dimensional LLM-judge rubric. a response passes when the format is valid and every 1–4 dimension scores above 2. conditional quality in all tables is the average of the four 1–4 scores over responses whose target action is correct. the judge prompt is versioned and will be released with the code.

Dimension	Scale	Measures
Format	0/1	valid JSON with answer and image-evidence fields
Correctness	1–4	factual accuracy of the answer
Grounding	1–4	claims supported by cited evidence
Brevity	1–4	conciseness, no redundancy
Helpfulness	1–4	usefulness for the user's goal

Table 12: Reward components actually passed to GRPO for the two reward conditions, as recorded in the frozen launcher and the preregistration. R_0 is the judge-only configuration; R_1 adds the certificate-routed components. **Scale** is the range of the component value *before* weighting. The gate component is registered and logged but carries weight 0.00 in R_1, so it contributes nothing to the optimized return and is a diagnostic only — it is not a hard gate during training. Component weights sum to 0.32 (R_0) and 0.92 (R_1); no renormalization is applied. If the judge response cannot be parsed, all four judge dimensions are set to 0 for that completion; if a training prompt has no certificate entry, action, factual and gate are set to 0 (neutral). D_1 has certificate coverage $400/400$.

Component	Source	Scale	R_0	R_1
`format`	Local rule (JSON schema)	$[0, 1]$	0.02	0.02
`correctness`	LLM judge, rubric v5	$(s-1)/3$ from 1–4	0.03	0.03
`grounding`	LLM judge, rubric v5	$(s-1)/3$ from 1–4	0.25	0.25
`brevity`	LLM judge, rubric v5	$(s-1)/3$ from 1–4	0.01	0.01
`helpfulness`	LLM judge, rubric v5	$(s-1)/3$ from 1–4	0.01	0.01
`action`	Certificate L4 target vs. $\hat{a}$	$\{+1, -1\}$	—	0.30
`factual`	Certificate verifier	$[-1, 1]$	—	0.30
`gate`	Hard-violation indicator	$\{-1, 0\}$	—	0.00
Total weight			*0.32*	*0.92*

qualifiers are checked for non-abstaining responses whose target is **Answer**. Over-refusal requires a target of **Answer**, a realized action of **Abstain**, and at least one hard evidence atom. All 100 target-**Answer** evaluation examples satisfy this evidence requirement.

The factual score follows Eq. 3, with $\alpha_u = 0.2$, $\alpha_h = \alpha_f = 0.5$, $\alpha_{\text{or}} = 1.0$, and $\alpha_q = 0.1$, and is clipped to $[-1, 1]$. A score of 1 indicates that no certificate-scoped deduction was triggered.

Reward configurations. R_0 uses five quality components with weights 0.02 for format, 0.03 for correctness, 0.25 for grounding, 0.01 for brevity, and 0.01 for helpfulness. R_1 retains these components and adds action and factual rewards with $\lambda_a = \lambda_f = 0.30$. Weights are not renormalized. The severe-violation channel is logged with $\lambda_g = 0$ and does not contribute to training.

The action reward is $+1$ when $\hat{a} = a^\star$ and -1 otherwise. The factual reward is the verifier score. These certificate-based components are optimized in the six R_1 runs. Each training completion receives one judge evaluation at temperature 0.1 with a 1,000-token budget. Table 12 summarizes the configurations.

Metrics and aggregation. We report action accuracy, four-class macro-F1, over-refusal, under-refusal, factual score, severe-violation rate, and strict JSON compliance. Under-refusal denotes an **Answer** response when the target is **Ask**, **Abstain**, or **Escalate**. Both refusal rates use all valid responses as the denominator. The severe-violation rate counts responses triggering over-refusal, a hard-negative match, or a forbidden claim. Strict JSON requires exactly `reasoning`, `answer`, and `image_evidence`, all with string values.

Table 13: Evaluation of multimodal models on the 400-item EviVQA test set. GRPO results are arithmetic means across three training seeds; other models are evaluated once. Rates and macro-F1 are percentages, factual scores use the $[-1, 1]$ scale, and response-quality scores use the 1–4 scale. GPT-5.4 (OpenAI, 2026) and Claude Sonnet 4.6 (Anthropic, 2026) results exclude 4 and 1 API-failed items, respectively. Gemini-3.8-Flash (Google DeepMind, 2026) is evaluated on a non-stratified 40-item subset and reported for descriptive reference only. Response quality is evaluated by `qwen3-vl-plus`; JSON denotes strict schema compliance. Dashes indicate that full-test results are unavailable.

Model	**Action Selection**		**Refusal Behavior**		**Certificate Checks**		**Format**	**Response Quality**			
	Acc.↑	Macro-F1↑	Over-ref.↓	Under-ref.↓	Factual↑	Gate Viol.↓	JSON↑	Correct.↑	Ground.↑	Brevity↑	Helpful.↑
General-Purpose Multimodal Models											
GPT-5.4 (OpenAI, 2026)	44.19	37.08	0.00	42.17	0.867	9.34	100.00	3.93	3.99	3.99	3.99
Claude Sonnet 4.6 (Anthropic, 2026)	32.08	21.63	0.00	59.15	0.864	8.27	45.11	3.88	3.96	3.97	3.98
Gemini-3.8-Flash (Google DeepMind, 2026)	27.50	20.4	0.00	72.50	0.860	10.00	5.00	3.42	3.58	3.60	3.58
Kimi-K3 (Kimi Team et al., 2026)	39.25	33.09	0.25	50.25	0.877	7.50	96.75	3.84	3.90	3.90	3.90
Qwen3.8-Max (Qwen Team, 2026b)	40.75	32.05	0.25	47.50	0.868	7.50	95.75	3.97	3.98	3.99	4.00
Qwen3.8-Omni-Flash (Qwen Team, 2026c)	46.00	39.54	1.00	37.25	0.874	8.25	99.75	3.92	3.99	4.00	4.00
Qwen3.7-Plus (Qwen Team, 2026a)	40.50	34.07	0.00	47.50	0.877	8.00	95.75	3.87	3.97	4.00	3.95
Backbone and Supervised Baseline											
Qwen3-VL-4B-Instruct (Base) (Bai et al., 2025)	45.75	39.94	0.25	40.00	0.911	4.00	100.00	3.25	3.78	4.00	3.56
SFT	46.00	40.88	0.50	42.75	0.877	7.75	100.00	3.66	3.95	3.98	3.88
GRPO-Trained Models (Shao et al., 2024)											
GRPO (D_0R_0)	46.25	41.04	0.42	42.58	0.882	7.00	100.00	3.662	3.942	3.975	3.867
Ours (D_1R_1)	47.50	42.55	0.67	41.67	0.880	7.58	100.00	3.654	3.947	3.968	3.867

Configuration means average three seed-level results; macro-F1 is computed separately for each seed before averaging. Judge dimensions are averaged over scored responses. Statistical analyses preserve example-level pairing across configurations.

C.3 Additional Results and Statistical Analysis

Detailed model comparisons. In Table 13, D_1R_1 achieves 47.50% action accuracy and 42.55% macro-F1, compared with 45.75%/39.94% for Base and 46.00%/40.88% for SFT. The fully evaluated general-purpose models obtain 32.08–46.00% action accuracy, despite generally higher correctness and helpfulness scores. Under-refusal is the dominant error: D_1R_1 answers 65.5% of evaluated requests, although only 25% have **Answer** as the target.

Factorial effects. Table 4 reports the four configuration means and factorial contrasts. Relative to D_0R_0, D_1R_1 improves action accuracy by 1.25 percentage points and macro-F1 by 1.51 points, while reducing under-refusal from 42.58% to 41.67%. Over-refusal increases from 0.42% to 0.67%, and severe violations from 7.00% to 7.58%; the factual score changes from 0.882 to 0.880. The reward factor contributes more to action accuracy than the data factor (+1.00 versus +0.25 points), with an interaction of +0.33 points.

Paired uncertainty estimates. We use a paired hierarchical bootstrap with 10,000 draws and random seed 20260924. Each draw resamples training seeds and evaluation examples while preserving the four-configuration pairing. We report percentile 95% confidence intervals and apply Holm correction across the seven metrics within each contrast.

For $D_1D_1 - D_0R_0$, the action-accuracy difference is $+1.25$ points (95% CI $[-1.42, 4.00]$), the under-refusal difference is -0.92 points ($[-4.00, 2.00]$), and the factual-score difference is -0.0022 ($[-0.0146, 0.0102]$). None of the corrected contrasts reaches statistical significance.

Common-intersection analysis. The ten full-set model rows share 395 examples with successful responses, excluding four GPT-5-4.1 API failures and one Claude Sonnet 4.1 failure. On this common set, D_1R_1 obtains 47.68% action accuracy and 42.51% macro-F1. Base retains the highest factual score (0.910) and the lowest severe-violation rate (4.05%). The leading point-estimate rankings remain unchanged after aligning the evaluation denominators.

Supplementary mixed-effects analysis. We fit binomial logistic mixed-effects models to action correctness, over-refusal, under-refusal, and severe violations. Fixed effects include data, reward, their interaction, and target action; random intercepts account for example and run. The models are fitted by variational Bayes. For action accuracy, the reward coefficient is 0.154 with posterior standard deviation 0.090. The example-level random-intercept standard deviation is 3.58, compared with 0.141 at the run level, indicating greater variation across examples than across runs.

C.4 Reproducibility and Scope

Offline recomputation. The reviewer archive includes scoring scripts, saved predictions, judge scores, and analysis outputs. CPU-only recomputation reproduces the main metrics, factorial results, common-intersection results, bootstrap intervals, and generated tables and figures. The supplementary mixed-effects fit is retained from the original analysis; refitting requires `statsmodels` and `patsy`.

Resource requirements. Training uses four GPUs per run, with two concurrent runs on an eight-GPU machine. The 12 runs generate 38,400 training completions and require the same number of judge calls before retries. Regenerating predictions and judge scores requires the corresponding model weights, image inputs, and model-endpoint access.

Anonymization. The archive replaces identifying paths and hostnames with placeholders. URL anonymization changes the displayed Claude Sonnet 4.6 factual score from 0.864 to 0.865. All reported tables use the anonymized value; the GRPO results are unchanged.