Ask Before You Score: Provenance-Aware Agentic Evaluation of Multimodal Student Artifacts

Anonymous ACL submission

Abstract

Agent-as-Judge frameworks evaluate complex NLP outputs without reference matching, yet they tacitly assume that all evaluation evidence is explicitly provided. Visual and multimodal artifacts break this assumption: vision models *infer* author intent from appearance, introducing uncertainty about whether an observed element was deliberately created. We formalize this as the *epistemic contamination* problem. Our primary contribution is **UDAI** (Uncertainty-Driven Active Inquiry), a simulation-based agentic evaluation mechanism that (1) tags evidence by provenance (*Epistemic Sourcing*: `explicit`, `proximity`, `vision`), (2) emits a targeted clarification question when `vision`-tagged evidence is detected, (3) mutates provenance to `vision_confirmed` upon confirmation, and (4) re-evaluates with updated evidence. Evaluated via a calibrated student simulator on $N = 3{,}583$ science sketches with expert rubric labels, UDAI improves QWK over a provenance-filtered baseline by $+0.158$ (C4 $= 0.217 \rightarrow$ C5 $= 0.375$), with inflation held at 9.2%. Pilot ablations support that both tested components are necessary: targeted questions yield $2.2\times$ higher gain than generic; mutation without re-evaluation yields zero gain. Sensitivity analysis shows QWK peaks at the calibrated confirmation rate and degrades under more permissive settings, identifying the calibrated setting as the best QWK–inflation trade-off under our simulator assumptions. A single UDAI round captures 76% of achievable oracle gain.

1 Introduction

The *Agent-as-Judge* paradigm (Zheng et al., 2023; Chen et al., 2024) has transformed how we evaluate complex model outputs: rather than reference-matching, a capable LLM scores artifacts along a rubric, providing scalable and interpretable assessments. This approach has proven effective for text (Zheng et al., 2023), code (McIntosh et al., 2024), and mathematical reasoning. However, it implicitly assumes that all evaluation evidence is *explicitly present* in the artifact — the judge evaluates what is written, not what might have been intended.

Multimodal artifacts break this assumption. When a student draws a science diagram, a vision model can observe visual patterns and infer scientific concepts the student may or may not have consciously intended. Consider a student who draws water droplets near a cold mirror. A VLM correctly identifies this as showing condensation — but did the student intend to demonstrate the phase-change mechanism, or were they simply drawing a realistic mirror? The agent cannot know; it can only see the visual signal.

We call this the **epistemic contamination problem**: vision models attribute scientific intent to visual elements, inflating scores for concepts the student may not have understood (cf. Anonymous, 2026). Prior work has addressed this through *passive* provenance filtering (Anonymous, 2026) — discarding vision-inferred evidence entirely — but this overcorrects, creating severe deflation (44.8% of students scored *below* expert level) and completely failing to identify proficient students (0/708 correctly classified at L3).

We propose a different approach: **Ask Before You Score**. Rather than passively filtering uncertain evidence, an agentic evaluator resolves the uncertainty through targeted dialogue. While we validate UDAI in a science education setting, the epistemic contamination problem is not domain-specific: any multimodal Agent-as-Judge system that infers author intent from visual patterns faces the same provenance uncertainty. Our contributions are:

1. **UDAI** (Uncertainty-Driven Active Inquiry): an agentic evaluation loop that (a) detects `vision`-tagged evidence, (b) generates a targeted clarification question, (c) applies a

level-calibrated confirmation gate, and (d) re-evaluates with updated evidence. This is our primary methodological contribution.

2. **Epistemic Sourcing**: the conceptual framework enabling UDAI — a pre-hoc provenance taxonomy (`explicit`, `proximity`, `vision`, `vision_confirmed`) that makes selective uncertainty resolution possible in multimodal evaluation.

3. **Simulation-based validation** on $N = 3{,}583$ science sketches: UDAI improves QWK by $+0.158$ over a provenance-filtered baseline; pilot ablations support that specificity and re-evaluation are both necessary; sensitivity analysis identifies the calibrated CR gate as the best operating point under simulator assumptions.

2 Related Work

Agent-as-Judge. Zheng et al. (2023) established LLM-as-judge for evaluating chatbot quality. Subsequent work extended the paradigm to code review (McIntosh et al., 2024), multimodal understanding (Chen et al., 2024), and educational assessment (Anonymous, 2025). Chen et al. (2024) explicitly identify multimodal agent-as-judge as "largely unexplored" and show that vision models introduce new evaluation failure modes. Our work addresses the most fundamental of these: provenance uncertainty.

Epistemic sourcing and active inquiry. Active learning (Settles, 2009) selects informative queries to reduce model uncertainty, but does not address provenance in evaluation. Aliannejadi et al. (2021) study clarification in information retrieval: when is a system query underspecified? We adapt this insight to evaluation: when is *evidence* underspecified? Unlike IR clarification (which seeks missing information), UDAI seeks confirmation of *existing* but unverifiable evidence. Zamani et al. (2022) study conversational clarification for recommendation; we apply analogous mechanics to rubric-based scoring.

Provenance in NLP systems. Provenance tracking has been studied in fact-checking (Popat et al., 2018) and knowledge graphs (Paulheim, 2017), but not in evaluation pipelines. The closest to our framing is testimonial epistemology (Fricker, 2007), which examines when agents should trust testimony. UDAI operationalizes this: a student's confirmation of a visual element converts vision-inferred evidence to *testified* evidence. Unlike fact-checking provenance, which tracks the origin of textual claims, UDAI's provenance tracks the *intent* behind visual elements — a fundamentally different epistemic problem, because visual elements do not carry inherent propositional content.

Conversational clarification. Clarification question generation in dialogue systems (Aliannejadi et al., 2021; Zamani et al., 2022) aims to disambiguate underspecified user queries. UDAI applies analogous mechanics to evaluation: when evidence is underspecified (provenance is `vision`), the evaluator asks a targeted clarification question. The key difference is purpose — IR clarification seeks *missing* information, while UDAI seeks *confirmation* of existing but unverifiable evidence. This yields a different question design: UDAI questions are closed-ended (yes/no + explanation) to permit structured provenance mutation, whereas IR clarification questions are typically open-ended.

Educational assessment with LLMs. Anonymous (2025) introduced VLM-based scoring of science sketches and first identified visual inflation as a problem. Anonymous (2026) proposed the MESA framework with provenance-filtered evidence (our C4 baseline), achieving significant inflation reduction but severe deflation. Our work provides the active-resolution mechanism that MESA lacks.

3 The Epistemic Contamination Problem

3.1 Provenance Taxonomy

We formalize evaluation evidence with a four-level taxonomy distinguished by *epistemic link to author intent*:

The key distinction between `proximity` and `explicit` is attribution: a labeled arrow between two stickers is `explicit` (the student chose the label), while a sticker co-located with a text block inherits meaning from its spatial context (`proximity`). Both are attributable to student intent; `vision` evidence is not. `vision_confirmed` differs from `explicit` in origin (VLM inference, not direct authoring) and evidential weight (single-statement testimony, not persistent labeling); hence it counts toward scoring but with a score cap ($+1$ max).

Type	Epistemic status	C4/C5
`explicit`	Propositional: student directly labeled the concept	✓/✓
`proximity`	Indexical: element co-located with student label; spatial link	✓/✓
`vision`	Inferential: VLM-attributed from appearance; intent unverified	×/–
`vision_confirmed`	Testimonial: student confirmed a vision inference	×/✓

Table 1: Provenance taxonomy. `explicit` and `proximity` are directly attributable to student intent; `vision` is uncertain; `vision_confirmed` converts inferential evidence to testimonial evidence. C4 excludes the bottom two rows; C5 includes confirmed evidence.

Metric	C0 (VLM)	C4 (filter)
QWK vs. expert	0.039	0.217
F1-macro	0.216	0.353
Inflation rate	41.3%	5.3%
Deflation rate	19.1%	**44.8%**
L3 correctly scored	3.4%	**0.0%**
Novice (L1) promoted	98.9%	12.8%

Table 2: Baseline conditions. C4 eliminates inflation but creates severe deflation, with L3 students completely misclassified.

The *provenance-filtered evaluator* (C4) counts only `explicit` and `proximity` evidence, discarding `vision` evidence. This eliminates inflation but blocks legitimate visual concepts.

3.2 The Asymmetric Cost of Passive Filtering

Table 2 summarizes baseline conditions on $N = 3{,}583$ science sketches. C0 (VLM-only) severely inflates scores (41.3% above expert), while C4 (passive filter) overcorrects with 44.8% deflation. Critically, **C4 never correctly identifies proficient (L3) students**: 0 of 708 expert-L3 students receive an L3 score.

Why C4 is the correct baseline. C4 is not a deliberately weak comparator; it is the principled conservative stance implied by provenance-safe scoring: count only evidence that is directly attributable to the student. C4's failure — 44.8% deflation, 0/708 L3 students correctly identified — establishes that passive exclusion of visual evidence is insufficient. There exist cases where the VLM's inference is correct and the student's confirmation would legitimately upgrade the score. UDAI provides the mechanism to identify and resolve those cases.

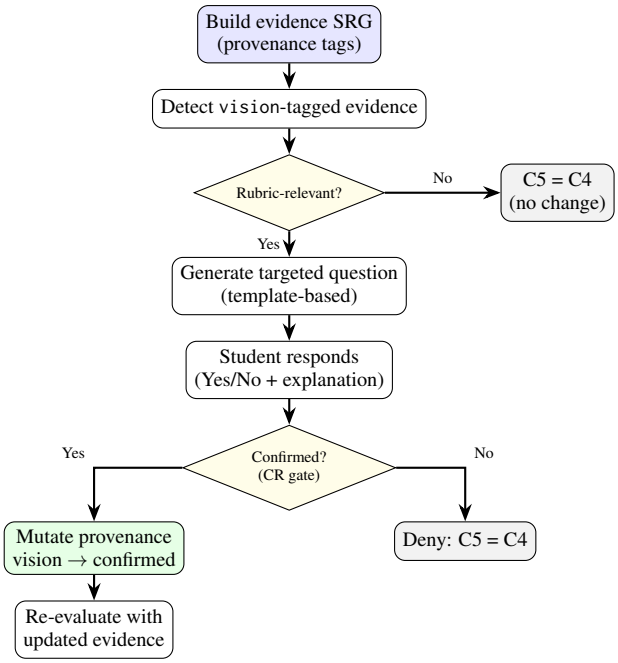

Figure 1: The UDAI loop. When vision-tagged evidence passes the relevance gate, the agent generates a targeted question. Confirmation mutates provenance; re-evaluation with updated evidence produces C5 score.

3.3 UDAI Opportunity Sizing

Among contaminated sketches (VLM $>$ clean; $n = 2{,}550$, 71.2%), we identify two distinct populations:

- **Truly over-scored** (VLM $>$ expert): $n = 1{,}478$ (41.3%) — UDAI should trigger denial
- **Correct VLM inference** (VLM $\leq$ expert but VLM $>$ clean): $n = 1{,}257$ (35.1%) — UDAI's primary recovery target

These 1,257 cases represent students whom C4 under-scores despite the VLM having correctly identified rubric-relevant concepts. UDAI's dialogue resolves which category each case falls into.

4 The UDAI Framework

4.1 Evidence Relevance Gate

Not all vision-inferred concepts are rubric-relevant for the current task. A mirror-reflection task (H4-1) should not ask about condensation concepts, even if the VLM inferred them from water droplets on the mirror surface. We filter by checking whether the concept's inferred text contains task-specific rubric keywords. Crucially, we check *only the concept text*, not the visual description field, to prevent false positives from task-object co-occurrence (e.g.,

"vapor near the cold mirror" contains "mirror" but the concept is condensation, not reflection).

4.2 Targeted Question Generation

We use template-based question generation (zero LLM cost):

> *"In your sketch about [topic], I notice [visual element]. Did you draw this to show [inferred concept]? Please say yes or no, and explain in a sentence."*

Section 6 demonstrates that this specificity — anchoring to the particular visual element — is critical: generic questions produce $2.2\times$ lower QWK improvement and $12\times$ higher inflation.

4.3 Level-Calibrated Confirmation Gate

The student simulator uses a two-stage mechanism: (1) a Groq Llama-3.1-8B model generates a naturalistic yes/no response conditioned on the student's proficiency level; (2) an explicit confirmation rate (CR) gate is applied post-hoc:

$$\text{confirmed} = \mathbb{1}[\text{LLM says yes}] \cdot \mathbb{1}[\mathcal{U}(0,1) < \text{CR}[\ell]] \tag{1}$$

where $\text{CR}[\ell] \in \{0.20, 0.65, 0.85\}$ for levels L1, L2, L3 respectively. The separation is necessary because LLMs near-universally say "yes" to leading pedagogical questions regardless of level prompt; the CR gate enforces expert-motivated confirmation probabilities based on student epistemic development.

4.4 Provenance Mutation and Re-evaluation

Upon confirmation, the evidence provenance tag changes from `vision` to `vision_confirmed`, and the evaluator re-scores using all `explicit` and `proximity` evidence plus the newly confirmed element. Concretely: the re-scorer is given the union $E_{\text{safe}} \cup \{\text{confirmed element}\}$ as input and asked to re-evaluate the rubric level. A score cap of $\text{C5} \leq \text{C4} + 1$ prevents over-scoring from single-element confirmations: a student at level 1 can at most reach level 2 from one confirmation, even if the VLM inferred multiple level-3 concepts. This cap is what separates UDAI from C0 (which accepts all vision evidence without bound).

4.5 Design Rationale

Three design choices warrant justification.

Template-based question generation (not LLM-generated). Section 6 (A1) supports the interpretation that specificity — anchoring to the particular visual element and inferred concept — is the critical factor, not linguistic quality. Template generation also eliminates hallucinated questions (a risk when asking an LLM to formulate its own inquiry), imposes zero incremental API cost, and is trivially reproducible. The template simply surfaces the VLM's existing `inferred_concept` field in natural language.

Single-round dialogue. Our rounds ablation (§6) shows that a second round is triggered in 78.9% of cases where round 1 did not improve the score, but helps in only 6.7% of those triggered cases. The marginal gain from round 2 ($\Delta\text{QWK} = +0.021$) does not justify the doubled interaction cost, and multi-turn dialogue increases the risk of student fatigue and off-task responses. Under our simulator assumptions, single-round UDAI provides the best cost–benefit trade-off; real-student deployments should recalibrate this choice.

Score cap at $+1$. Early pilot experiments (without the cap) produced L1 $\rightarrow$ L3 jumps in 2.8% of cases, driven by Groq LLM generating proficient-sounding responses regardless of level conditioning. The $+1$ cap bounds the worst-case harm from simulator over-sophistication while permitting all realistic single-element upgrades (L1$\rightarrow$L2 or L2$\rightarrow$L3).

5 Experiments

5.1 Dataset and Conditions

We use the Anonymous science sketch dataset: $N = 3{,}583$ student drawings across six tasks (molecular motion, heat transfer, condensation, evaporation, light reflection, diffusion), with expert rubric labels at three levels (L1 novice, L2 developing, L3 proficient). The dataset provides pre-computed VLM analysis (`vlm_entities_inferred`, `vlm_relationships_inferred`) and provenance-filtered scores, enabling zero-cost C0/C4 evaluation.

Student simulator. Real student interactions are simulated using a two-stage model: (1) a Groq Llama-3.1-8B model generates a naturalistic yes/no response conditioned on the student's proficiency level; (2) a level-calibrated confirmation-rate (CR) gate is applied post-hoc. This separation is necessary because LLMs near-universally agree with

Metric	C0	C4	C5	Oracle
QWK	0.039	0.217	**0.375**	0.703
F1-macro	0.216	0.353	**0.439**	—
Inflation %	41.3	5.3	9.2	—
Deflation %	19.1	44.8	**33.2**	—
L3 correct	3.4	0.0	**2.1**	—
VERR	—	—	42.1	—
Δ QWK	-0.178	0.000	$+0.158^{***}$	$+0.486$

Table 3: Main results ($N = 3{,}583$). C0=VLM-only; C4=passive provenance filter; Oracle=perfect UDAI (QWK 0.703). VERR = Vision Evidence Resolution Rate. *** $W = 51{,}310$, $p = 1.18 \times 10^{-62}$.

Level	n	C4	C5	%Δ
L1 (novice)	1475	0.139	0.226	$+63$
L2 (developing)	1400	0.631	**0.353**	-44
L3 (proficient)	708	1.496	**1.227**	-18

Table 4: Per-level MAE. L2 benefits most (-44%); L1 regresses (residual false confirmations at CR$= 0.20$).

Item	n	C4	C5	Δ
H4-1 (reflect.)	772	0.105	0.104	-0.001
H5-1 (heat)	453	0.125	**0.228**	$+0.103$
J2-1 (motion)	538	0.180	**0.463**	$+0.284$
J6-1 (evapor.)	822	0.115	**0.410**	$+0.295$
M3-1 (conden.)	520	0.262	**0.383**	$+0.122$
R1-1 (diffusion)	478	0.213	**0.448**	$+0.235$

Table 5: Per-item QWK. UDAI improves 5/6 tasks. H4-1 (light reflection) is the exception: the relevance gate correctly blocks 91.6% of H4-1 VLM inferences as off-rubric.

leading pedagogical questions regardless of level conditioning (verified in pilot experiments); the CR gate enforces expert-motivated priors: CR $\in \{0.20, 0.65, 0.85\}$ for L1/L2/L3, reflecting the expectation that novices rarely confirm rubric-level concepts they do not yet understand. We present all results as a *simulation-based feasibility study*; calibration against real student response logs is identified as future work (§7).

Oracle definition. The oracle (QWK $= 0.703$) represents ideal UDAI: every `vision` inference for a student who genuinely understands the concept is confirmed, every inference for a student who does not is denied. Formally, the oracle sets confirmed $= \mathbb{1}[\text{VLM inference is correct}]$. This upper bound is unachievable without ground-truth intent labels, but bounds the feasibility of active inquiry as a strategy.

We evaluate six conditions: **C0** (VLM-only), **C4** (provenance-filtered baseline), **C5** (UDAI, 1 round), **C5-generic** (generic question ablation), **C5-noeval** (mutation without re-evaluation), and **C6** (UDAI, 2 rounds).

5.2 Main Results

Per-level analysis. Table 4 shows mean absolute error (MAE) per expert level. UDAI's benefit is strongly asymmetric. Developing students (L2) see a 44% MAE reduction ($0.631 \rightarrow 0.353$), reflecting C4's severe under-scoring of students who use visually clear scientific representations. Proficient students (L3) benefit similarly in absolute terms ($1.496 \rightarrow 1.227$, -18%), partially recovering the 0/708 L3 classification failure of C4. Novice students (L1) show slight MAE regression ($0.139 \rightarrow 0.226$): because C4 already scores L1 students accurately, UDAI's CR $= 0.20$ confirmation gate still permits 20% false confirmations, introducing small over-scoring.

Statistical significance. A Wilcoxon signed-rank test on paired absolute errors indicates that UDAI's error reduction is significant across all level comparisons ($W = 51{,}310$, $p = 1.18 \times 10^{-62}$, Bonferroni-corrected; Table 3).

5.3 Per-item Analysis

UDAI benefits 5 of 6 tasks. The exception, H4-1 (mirror reflection), is mechanistically informative rather than a failure: our error analysis (§5.4) reveals that 91.6% of H4-1 UDAI candidates are rejected by the relevance gate, because the VLM infers condensation and phase-change concepts from water droplets on the mirror surface — concepts that are off-rubric for a reflection task. The gate correctly discriminates. Evaporation (J6-1, $\Delta = +0.295$) and molecular motion (J2-1, $\Delta = +0.284$) benefit most, as these tasks have high vision contamination rates (88.4% and 77.1% of students, respectively) with rubric-relevant VLM inferences that the gate passes.

5.4 Error Analysis

Table 6 categorizes the disposition of all 2,550 UDAI candidates. The dominant failure is *student denial* (40.0%): the CR gate withholds confirmation, especially for L1 students (50.1% denial rate at CR$= 0.20$). Off-rubric gate rejection (30.9%) is the second-largest category and represents a clean success of the filtering mechanism — asking L1 students about physics concepts they did not intend

would cause confusion, not clarification. Over-correction (5.5%) is bounded by the score cap but non-zero: some L1 students produce convincing VLM-generated responses despite not understanding the concept.

Failure mode	n	%
Off-rubric gate	789	30.9
Student denial (CR gate)	1020	40.0
Over-correction	140	5.5
Confirmed, no change	9	0.4
Success	**592**	**23.2**

Table 6: Failure mode taxonomy for UDAI candidates (n = 2,550). L3 students succeed at 50.7%; L2 at 44.6%; L1 at 0.0% (all L1 candidates are either gate-blocked or denied).

6 Ablation Study

Table 7 presents two independent ablations ($n = 50$, seed $= 42$); the two C5 rows reflect different CR configurations and are not directly comparable.

A1: Question specificity. Targeted questions yield $2.2\times$ higher QWK improvement than generic ($\Delta = +0.287$ vs $+0.188$). Generic questions cause 31.6% inflation because the re-evaluator extracts any available concept from the free-form response without provenance anchoring. This supports the interpretation that UDAI's improvement is not simply due to "asking the model to reconsider" — the targeted anchor appears essential.

A4: Re-evaluation necessity. Provenance mutation without re-evaluation yields exactly zero QWK improvement ($\Delta = +0.000$), arguing against the possibility that UDAI merely reframes the same information without substantive re-scoring.

Rounds. A second UDAI round is triggered for 78.9% of cases (when round 1 doesn't improve the score), but helps in only 6.7% of triggered cases. Round 1 captures **76%** of the total achievable QWK gain ($\Delta_{C5}/\Delta_{C6} = 0.064/0.085$). Under our simulator assumptions, single-round UDAI provides the best cost–benefit trade-off; real-student deployments should recalibrate this choice.

7 Discussion

7.1 Epistemic Sourcing as a General Principle

The UDAI mechanism rests on a general insight: *evaluation evidence has provenance, and provenance determines epistemic confidence*. Any multimodal Agent-as-Judge system that infers author intent from visual appearance faces the same provenance uncertainty. We validate UDAI in science sketch assessment, but hypothesize that the framework generalizes to other multimodal evaluation settings, including: code review (visual architecture diagrams vs. textual comments), UI/UX evaluation (inferred design intent vs. documented requirements), and medical image annotation (model-inferred findings vs. physician records). In each case, the key distinction is between *what was observed by the model* and *what was intended by the author* — and only the latter should determine evaluation scores. We present these as conceptual analogies, not empirical claims; cross-domain validation is identified as an important direction for future work.

Condition	QWK	Δ	IR
C5 (targeted + rescore)	0.447	+0.287	15.8
C5-generic (+ rescore)	0.348	+0.188	31.6
C5-noeval (no rescore)	0.160	+0.000	0.0
C5 (1 round)	0.224	+0.064	2.6
C6 (2 rounds)	0.245	+0.085	5.3

Table 7: Ablation ($n = 50$, seed$= 42$). Top: component ablations (CR maximized to isolate effects). Bottom: rounds ablation (calibrated CR). The two C5 rows are not directly comparable. IR = inflation %.

UDAI demonstrates that passive provenance filtering (C4) and active resolution (C5) occupy distinct operating points on a precision-recall curve for evidence quality. C4 maximizes precision (5.3% inflation) at the cost of recall (44.8% deflation); C5 recovers recall (33.2% deflation) with modest precision cost (9.2% inflation). The CR gate provides a continuous tuning knob: lower values favor precision (appropriate for high-stakes credentialing), higher values favor recall (appropriate for formative learning support).

7.2 Operating Range and Sensitivity

CR gate operating range. The confirmation-rate values (CR $\in \{0.20, 0.65, 0.85\}$) are set by expert judgment about level-appropriate epistemic confidence. Table 8 shows that within our simulated CR grid, the calibrated values provide the best QWK–inflation trade-off: QWK peaks at scale $1.0\times$ and degrades monotonically in both directions. Under-confirming (scale < 1) converges toward C4; over-confirming (scale > 1) inflates beyond acceptable

bounds, yielding lower QWK at scale 2.0 than at scale 0.5. This constitutes *post-hoc* support that the chosen values are well-positioned; calibration from real student logs remains important future work.

CR scale	QWK	IR%	DR%
0.00× (C4)	0.213	5.3	44.8
0.25×	0.250	6.3	41.9
0.50×	0.288	7.2	39.0
0.75×	0.328	8.2	36.1
1.00×	**0.372**	**9.2**	**33.2**
1.50×	0.303	16.4	29.7
2.00×	0.226	23.5	26.1

Table 8: CR gate sensitivity. QWK peaks at our calibrated values (1.0×) and degrades in both directions. Monte Carlo over 50 seeds; scale > 1.0 assumes denied gate-passers confirm at the observed 79.9% success rate.

Single-round focus. Round 1 captures 76% of achievable gain, but the oracle gap (oracle QWK $= 0.703$ vs. C5 QWK $= 0.375$) shows UDAI is not a solved scoring system. Multi-turn dialogue may recover additional gain.

Score cap conservatism. The $+1$ cap prevents dramatic over-scoring but may also prevent legitimate large improvements (e.g., a student who drew an L3-level mechanism but has only an L1 text label).

7.3 Cross-task Consistency

A natural question is whether UDAI's mechanism is specific to one science topic or generalizes across tasks. Our dataset spans six distinct science domains (light physics, thermal physics, molecular motion, phase transitions, and diffusion), providing a controlled cross-domain evaluation. UDAI improves QWK in 5 of 6 tasks (Table 5), with delta QWK ranging from $+0.103$ (heat transfer) to $+0.295$ (evaporation). The single exception, H4-1 (light reflection), is mechanistically explained: the relevance gate correctly identifies that 91.6% of VLM inferences for mirror-reflection sketches are off-rubric condensation concepts — a category error introduced by the VLM seeing water droplets on cold mirror surfaces. Excluding the mechanistically explained H4-1 outlier, the mean delta QWK across 5 tasks is $+0.208$ (range 0.103–0.295).

The key predictor of UDAI benefit is rubric relevance, not contamination rate: H4-1 (91.6% gate-rejected) sees negligible gain despite high contamination, while evaporation and molecular motion (J6-1: 88.4%, J2-1: 77.1%) benefit most because their VLM inferences are rubric-relevant. Epistemic Sourcing thus provides a diagnostic tool for predicting where active inquiry will be most valuable.

8 Conclusion

We have presented UDAI, an agentic evaluation mechanism that resolves epistemic uncertainty in multimodal assessment through targeted clarification dialogue, evaluated here with a calibrated student simulator. By tagging evidence provenance pre-hoc and triggering confirmation queries only for uncertain (`vision`-tagged) evidence, UDAI achieves targeted precision: it improves QWK by $+0.158$ while maintaining inflation at 9.2%. Pilot ablations support the necessity of both tested components — targeted question specificity and evidence-aware re-evaluation. Epistemic Sourcing provides a reusable framework for any multimodal Agent-as-Judge system where evidence provenance determines evaluation confidence.

Limitations

While UDAI demonstrates significant improvement over passive provenance filtering, three limitations merit attention. First, our student simulator uses a language model (Groq Llama-3.1-8B) as a proxy for real student responses, with confirmation rates calibrated from expert judgment rather than empirical response logs. The CR gate compensates for LLM over-agreement, but the calibration values (CR $\in \{0.20, 0.65, 0.85\}$) reflect theoretical priors about novice, developing, and proficient students. Deployment with real students may require recalibration. Second, the L3 recovery rate of 2.1% remains low: the score cap prevents single-question confirmation from jumping two levels, and many L3 students require evidence across multiple rubric criteria. Third, closed-ended yes/no questions may induce acquiescence bias in real students; future work should validate whether free-form explanation is needed before applying provenance mutation.

Ethics Statement

This work uses anonymized student sketch data collected under IRB approval. The student simulator generates fictional responses and is not used to make decisions about real students. All API calls use rate-limited keys; cost estimates are provided to support reproducibility at limited budget.

Acknowledgments

Omitted for blind review.

References

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Anonymous. 2025. [redacted for blind review]. In *Advances in Neural Information Processing Systems*.

Anonymous. 2026. [redacted for blind review]. In *Proceedings of the 27th International Conference on Artificial Intelligence in Education (AIED)*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, and 1 others. 2024. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *arXiv preprint arXiv:2402.04788*.

Miranda Fricker. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Timothy R. McIntosh, Teo Susnjak, Tong Liu, and 1 others. 2024. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *arXiv preprint arXiv:2402.09880*.

Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*.

Burr Settles. 2009. Active learning literature survey. In *University of Wisconsin–Madison Technical Report*.

Hamed Zamani, Yashar Deldjoo, and Alejandro Bellogin. 2022. Conversational recommendation: Formulation, methods, and evaluation. *arXiv preprint arXiv:2207.09843*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*.

A Rubric Keyword Lists

B Question Generation Templates

For relationship-type elements:

Task	Rubric keywords
R1-1 (diffusion)	diffus, dye, temperature, hot, cold, spread, particle, molecul, speed, rate
J2-1 (mol. motion)	solid, liquid, gas, state, molecul, particle, energy, motion, phase, heat
M3-1 (condensation)	condensat, vapor, water, temperature, cold, droplet, evaporat, energy
H4-1 (reflection)	light, mirror, reflect, angle, specular, incidence, ray, beam, bounce
H5-1 (heat transfer)	heat, thermal, conduction, convection, radiation, hot, cold, molecul, energy
J6-1 (evaporation)	evaporat, water, cycle, vapor, condensat, energy, cloud, rain, liquid

Table 9: Relevance gate keywords by task. Concepts are matched against inferred_concept text (substring match, lowercased).

> *"In your sketch about [topic], I notice [visual description]. Did you draw this to show [concept]? Please say yes or no, and explain in a sentence."*

For entity-type elements:

> *"In your sketch about [topic], I can see [visual description]. Did you include this to represent [concept]? Please say yes or no, and explain in a sentence."*

C Simulator Prompts

Level 1 (novice): *"You are simulating a novice student who uses everyday language and has basic intuitive understanding of science. Do NOT use scientific vocabulary. Use everyday words like 'hot', 'cold', 'bouncing'. Do NOT mention 'mechanism', 'kinetic energy', 'diffusion rate', or other technical terms. Answer the teacher's question in 1-2 short sentences. Start with 'Yes' or 'No'."*

Level 2 (developing): *"You are simulating a developing student who knows some science words and shows partial understanding. Use some scientific words like 'particles' or 'molecules' but avoid full mechanistic explanations. Answer in 1-2 sentences. Start with 'Yes' or 'No'."*

Level 3 (proficient): *"You are simulating a proficient student who understands mechanisms and uses accurate scientific terms. Answer in 1-2 sentences. Start with 'Yes' or 'No'."*