What You Score Matters More Than How You Score It: Criterion Type Outweighs Scoring Architecture in LLM Judge Reliability for Educational Assessment

Anonymous ACL submission

Abstract

LLM judges are increasingly deployed for rubric-based scoring of student responses, yet reliability varies enormously across criteria within a single task. Prior work attributes this variance to architectural choices—single-call prompting versus chain-of-thought, few-shot exemplars, or multi-agent decomposition—and proposes increasingly elaborate scoring pipelines. We challenge this assumption. In a systematic study of 13 scoring architectures applied to 44 rubric criteria across nine middle-school science tasks (572 architecture–criterion pairs, 50 student responses each), we find that architecture choice explains only $\eta^2 = 0.024$ (2.4%) of quadratic weighted kappa variance and is not a statistically significant predictor ($F = 1.14$, $p = 0.33$), whereas criterion *evidence type*—a five-category taxonomy of what students must demonstrate—explains $\eta^2 = 0.215$ (21.5%), **approximately 9× more** variance with far fewer levels (approximately 7.3× under partial η^2 from a joint two-factor model). Criteria requiring students to label an abstract concept or demonstrate reasoning are among the hardest (*concept_labeled* mean QWK= 0.428; *reasoning_shown* mean QWK= 0.296), yet counter-intuitively, LLMs *over-score* both types (*concept_labeled*: precision= 0.475, recall= 0.807; *reasoning_shown*: precision= 0.507, recall= 0.708) rather than under-score them. A two-phase evidence-type scoring method supports this diagnostic: it reduces over-scoring on concept criteria (+0.091 ΔQWK) while inadvertently tightening prompts for already-calibrated fact-stated criteria (−0.263 ΔQWK) and a single-call type-aware agent (A13) widens rather than closes the criterion-type QWK gap (η^2: 0.304 → 0.498). These results position the two-phase method as a *diagnostic instrument* that supports the characterisation of type-specific failure modes, not a universally superior scoring architecture. Large between-question QWK variance on SciEntsBank ($\sigma = 0.399$) confirms that strong item-level judge heterogeneity is not unique to our dataset, though the evidence-type ordering itself awaits full SciEntsBank annotation. We argue that rubric designers and evaluation practitioners should audit criterion evidence types before deploying an LLM judge, attending to criterion type first, scoring architecture second, and reserving local architecture tuning for within-criterion optimization on high-risk criteria.

1 Introduction

LLM judges are now routinely deployed for rubric-based scoring of student short answers in science and mathematics education (Latif and Zhai, 2024; Zhang et al., 2024), where manual grading at scale is cost-prohibitive and where timely formative feedback is known to improve learning outcomes. The attraction is clear: a single prompted call to a capable model can, in principle, replicate the scoring decision a teacher would spend minutes reaching—and do so for thousands of responses overnight. Yet this promise rests on an uncomfortable empirical fact: LLM judge reliability varies enormously across rubric criteria, even within a single assessment task and with a fixed model.

The problem practitioners face. Consider a representative middle-school science task with five rubric criteria. Using the same single-call prompting strategy and the same model, one criterion asking students to state a measured fact (*fact_stated*) achieves a quadratic weighted kappa (QWK) near 0.78, while a criterion asking them to label an abstract causal concept (*concept_labeled*) achieves QWK≈ 0.28—a gap of 0.50 QWK points on the same task, with the same students, scored by the same judge.

The natural response from the research community has been to propose increasingly elaborate scoring architectures: chain-of-thought prompting (Liu et al., 2023), reference comparison, few-shot exemplar selection (Kim et al., 2024), multi-agent critique, majority voting—each aimed at narrowing the performance gap by changing *how* the LLM is asked to score. But what if the bottleneck is not the scoring procedure, but the nature of the criterion being scored?

The gap in prior work. No prior study has systematically decomposed the variance in LLM judge performance and asked whether architecture selection is the right lever to pull. Research in LLM-as-judge evaluation has focused on documenting model-level biases—position effects, verbosity preferences, self-enhancement tendencies (Zheng et al., 2023)—and on calibrating aggregate judge reliability. Research in automated short-answer grading has benchmarked architectures and models on fixed task sets (Dzikovska et al., 2013; Sung et al., 2019), reporting task-level aggregate metrics that obscure within-task criterion heterogeneity. Neither line of work has asked: *what property of the criterion being scored determines whether an LLM judge can score it reliably?*

Our answer. Although our empirical setting is educational short-answer scoring, the underlying problem is general to rubric-based LLM evaluation: criteria differ in the communicative act they require, and these acts induce distinct judge failure modes. We conduct the first systematic two-factor study crossing 13 scoring architectures with 44 rubric criteria (nine middle-school science tasks, 50 responses per cell, 572 cells total). Architecture explains only $\eta^2 = 0.024$ (*non-significant*, $p = 0.33$), while a five-category taxonomy of *criterion evidence type* explains $\eta^2 = 0.215$ ($p < 10^{-29}$)—approximately 9× more. The dominant failure story is also inverted: *concept_labeled* criteria (precision= 0.475, recall= 0.807) and *reasoning_shown* criteria (precision= 0.507, recall= 0.708) are hard because LLMs *over-score* them, not because they are too strict. Within any single criterion the best architecture outperforms the worst by a median of 0.422 QWK points (Appendix B), confirming that architecture choice still matters locally even if it provides negligible global leverage.

Contributions.

- **Variance decomposition.** One-way ANOVAs on 572 cells: evidence type ($\eta^2 = 0.215$) explains ≈ 9× more QWK variance than architecture ($\eta^2 = 0.024$, non-significant); joint partial η^2 ratio is 7.3× ($\eta^2_{\text{partial}} = 0.220$ vs. 0.030).

- **Evidence-type taxonomy.** Five categories grounded in educational assessment theory (Anderson et al., 2001)—*fact_stated*, *data_compared*, *reasoning_shown*, *mechanism_explained*, *concept_labeled*—with an empirically characterized QWK ordering (0.607 to 0.296, gap= 0.311) and per-type precision–recall signatures.

- **The over-scoring finding.** Both *concept_labeled* (precision= 0.475, recall= 0.807) and *reasoning_shown* (precision= 0.507, recall= 0.708) are hard because LLMs are *too permissive*, inverting the prevailing assumption of overly-strict LLM judges.

- **Diagnostic method and external validation.** A two-phase evidence-type method confirms the taxonomy: +0.091 ΔQWK on *concept_labeled*, −0.263 on *fact_stated*; a single-call ablation (A13) widens rather than closes the type gap (η^2: 0.304 → 0.498). Large between-question variance on SciEntsBank ($\sigma = 0.399$, $F(21, 25) = 6.80$, $p < 0.0001$) confirms that strong item-level judge heterogeneity is not dataset-specific.

- **Cross-model robustness.** On two additional backbone models (gpt-4.1-mini, gpt-5.4-mini, $n = 26$ criteria), criterion-type η^2 remains ≈0.40; gpt-5.4-mini achieves *lower concept_labeled* QWK (0.337 vs. 0.443; Wilcoxon $p = 0.006$), challenging the hypothesis that scaling closes the criterion-type gap.

This work is complementary to selective-prediction approaches (Geifman and El-Yaniv, 2017; Angelopoulos et al., 2022), which address *when* an LLM judge should abstain; we address

what makes certain criteria harder in the first place.

2 Related Work

LLM-as-Judge. Zheng et al. (2023) introduced MT-Bench and showed GPT-4 achieves $> 80\%$ agreement with human preferences while documenting position-effect and verbosity biases. Liu et al. (2023) use probability-weighted scoring; Kim et al. (2024) fine-tune a dedicated open-source evaluator. None analyses criterion *type* as a predictor of per-criterion judge reliability; the variance we find across rubric criteria is invisible at the aggregate level these works report.

Criterion Decomposition in LLM Evaluation. Min et al. (2023) decompose long-form text into atomic facts; Lee et al. (2025) decompose criteria into yes/no checklists, improving inter-annotator agreement by 0.45 on average; Saad-Falcon et al. (2025) treat criteria as unit tests, surfacing 131% more error modes than holistic judges. These works show how decomposing criteria improves scoring. Our work is orthogonal: even after fixing the architecture, the *semantic type* of each criterion explains $9\times$ more QWK variance under marginal estimates, and characterising which types are inherently harder is a prerequisite for targeted decomposition.

Automated Short-Answer Grading. Dzikovska et al. (2013) introduced SciEntsBank (our external validation set, §7); Sung et al. (2019) and Camus and Filighera (2020) show domain-adaptive BERT improvements on these benchmarks; Latif and Zhai (2024) show fine-tuned ChatGPT narrows the gap to supervised baselines. These papers report aggregate metrics per task and do not ask whether criterion type is the primary driver of per-criterion performance variation.

Concurrent and Complementary Work. Chu et al. (2026) detect per-criterion FP/FN patterns from a labelled calibration set and apply targeted prompt repairs—post-hoc and response-conditioned, whereas our taxonomy is pre-deployment and can flag high-risk criteria before any student data is collected. Yu et al. (2025) auto-extract response-conditioned evaluation criteria at inference time (r=0.78); Hardy (2026) meta-analyse AI short-answer grading and find task–model difficulty misalignment as a primary failure mode, which our taxonomy makes structurally explicit. Wang et al. (2026) show a two-agent pipeline improves complex rubrics; our two-phase method shares the decomposition intuition but grounds phase selection in a theory-derived taxonomy.

3 Evidence-Type Taxonomy

3.1 Motivation

Rubric criteria differ not only in topic but in the *communicative act* they demand of the student. We propose that this distinction—what evidence a correct response must *contain*, regardless of topic—is more predictive of LLM judge behaviour than any architectural choice, and formalise it as a five-category *evidence-type taxonomy* derived from the Claim–Evidence–Reasoning (CER) framework used in science education.

3.2 The Five Evidence Types

Each type is defined by the *minimal communicative act* a student must perform to satisfy the criterion. Table 1 gives definitions and representative examples.

fact_stated. Assert a specific verifiable fact; narrow surface form (P= 0.911, R= 0.871).

data_compared. Compare ≥ 2 values or patterns from a stimulus; performance depends on whether reference values are rubric-explicit.

reasoning_shown. Connect evidence to a claim with an explicit logical link (CER chain).

mechanism_explained. Describe a causal process; LLMs demand precise mechanistic language (P= 0.923, R= 0.500; case study).

concept_labeled. Label or apply a scientific concept by category; hardest over-scoring type (P= 0.475, R= 0.807)—vanilla prompts award credit for any mention of the concept domain.

We compare against Bloom's cognitive level in Appendix F; evidence type subsumes Bloom's ($\Delta R^2 < 0.001$ when adding Bloom's to the evidence-type model).

3.3 Annotation Protocol

Two authors independently classified all 44 criteria using a keyword-anchored decision tree

Table 1: Evidence-type taxonomy: minimal communicative act per type.

Type	Definition	Example (crit. / response)
fact_stated	Assert a specific fact	*C*: Sugar types processed differently. *R*: "Not treated the same."
data_compared	Compare ≥2 stimulus values	*C*: Balloon A lost more mass than B. *R*: "A: 1.1 g, B: 0.3 g."
reasoning_shown	Claim + evidence + explicit link	*C*: Why more sugar ↑ fermentation. *R*: "More sugar → yeast food → CO_2."
mechanism_explained	Causal process (how/why)	*C*: How density forms liquid layers. *R*: "Denser molecules packed tighter."
concept_labeled	Label concept by name/category	*C*: Layers differ in density. *R*: "Oil floats; lower density."

("states that" → *fact_stated*; "explains why" → *reasoning_shown*; "identifies" / "labels" → *concept_labeled*; etc.). Post-calibration $\kappa = 0.880$ (40/44, 90.9%) after a joint calibration session; four residual disagreements were resolved by the decision tree as tiebreaker. Full protocol and per-criterion classifications appear in Appendix E.

4 Architecture Study

4.1 Experimental Setup

We evaluate 13 scoring architectures (A0–A12) across 44 binary rubric criteria from nine middle-school science tasks (50 student responses per criterion), using GPT-4o-mini as the backbone model, selected for its cost–accuracy balance on pilot experiments. Full architecture descriptions appear in Appendix A; the evaluation metric is quadratic weighted kappa (QWK), yielding $13 \times 44 = 572$ per-cell observations.

4.2 Architecture Rankings

Architectures cluster within a 0.103 QWK range (A0 single-call ranks 2nd at 0.485; A2 extract-then-judge is worst at 0.384 due to information loss). The full 13-row ranking appears in Appendix A; no architecture is a significant predictor ($F = 1.14$, $p = 0.33$).

4.3 Variance Decomposition

Table 2 and Figure 1 report the η^2 decomposition.

One-way (marginal) estimates. Architecture is *not* a statistically significant predictor of QWK ($F = 1.14$, $p = 0.33$), and the one-way marginal η^2 attributes only 2.4% of variance to architecture ($\eta^2 = 0.024$). A coarse five-way evidence-type classification yields a marginal $\eta^2 = 0.215$ ($F = 38.81$, $p < 10^{-29}$)—approximately 9× higher, with fewer than half as many levels. Because these are marginal estimates from separate models, they cannot be interpreted as non-overlapping variance components (their sum exceeds 1.0); the ratio is therefore an approximate comparison rather than a strict partition.

Joint model (partial η^2). A two-factor OLS model (`QWK ~ Architecture + EvidenceType`, Type II SS) yields partial $\eta^2 = 0.030$ for architecture ($p = 0.36$) and 0.220 for evidence type ($p \approx 0$), a **7.3**× ratio; the architecture × evidence-type interaction is non-significant ($F(48, 273) = 0.78$, $p = 0.85$).

Criterion identity (44 levels, $\eta^2 = 0.597$) is the dominant factor; task ($\eta^2 = 0.223$) accounts for more variance than architecture—*what* is scored matters far more than *how*.

The practical implication: attending to evidence type purchases 7.3× as much unique variance ($\eta^2_{\text{partial}} = 0.220$) as architecture search (partial $\eta^2 = 0.030$), and the classification can be made from rubric text alone before any LLM scoring is run.

Within any single criterion, the best architecture outperforms the worst by a median of 0.422 QWK points (full per-criterion breakdown in Appendix B), confirming that architecture choice still matters locally even if it provides negligible global leverage.

Robustness. A mixed-effects model with random intercepts for criterion and task confirms evidence type ($p < 10^{-6}$) and non-significant architecture ($p > 0.30$); see Appendix D.

Table 2: Variance decomposition ($n = 572$ cells). Marginal η^2 (one-way ANOVAs, sum > 1) and partial η^2 (joint two-factor OLS, Type II SS, non-overlapping). Evidence type explains ≈7.3× more unique variance than architecture.

Factor	Levels	Marginal η^2 (one-way)	Joint model (partial η^2) η^2_{partial}	F	p
Architecture	13	0.024	0.030	1.10	0.36
Evidence type	5	0.215	0.220	37.80	≈ 0
Criterion identity	44	0.597	—	—	—
Task	9	0.223	—	—	—
Sum of marginal η^2			1.058 (overlap confirmed)		

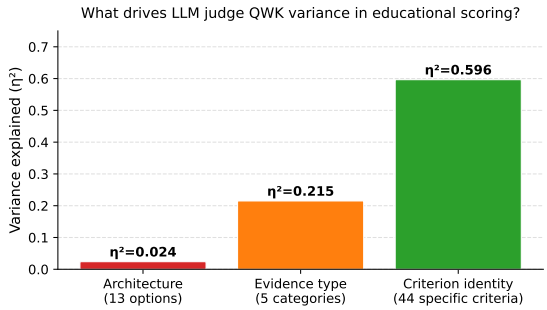

Figure 1: Marginal η^2 (one-way ANOVA) for each factor. Evidence type explains ≈9× more variance than architecture under marginal estimates (7.3× under partial η^2; Table 2). Architecture is not significant ($F = 1.14$, $p = 0.33$).

5 Evidence Type Outweighs Architecture

5.1 QWK by Evidence Type

Table 3 reports mean QWK for each evidence type using A0 as the reference architecture. The gap between *data_compared* (QWK= 0.607) and *reasoning_shown* (QWK= 0.296) is **0.311 QWK points**—more than 3× the architecture range of 0.103 (Kruskal-Wallis $H = 13.95$, $p = 0.007$; *mechanism_explained*, $n = 1$, is a case study only).

5.2 Precision–Recall Signatures by Type

Figure 2 plots every criterion in precision–recall space. *concept_labeled* and *reasoning_shown* are hard because LLMs *over-score* them: *concept_labeled* achieves recall= 0.807 but precision= 0.475; *reasoning_shown* ($n = 9$) shows precision= 0.507, recall= 0.708. LLMs credit any mention of the concept domain or partial reasoning structure—a failure of discernment, not sensitivity. *mechanism_explained* shows the opposite (precision= 0.923, recall= 0.500; case study). The type ordering is robust to architecture choice: the architecture × evidence-type interaction is non-significant ($F(48, 273) = 0.78$, $p = 0.85$).

Table 3: Mean QWK, precision, and recall by evidence type (A0, 44 criteria, $n = 50$ each). Gap: 0.311 QWK pts (*data_compared* to *reasoning_shown*; Kruskal-Wallis $H = 13.95$, $p = 0.007$; no pairwise contrasts survive Holm). †*mechanism_explained*: single criterion (case study only).

Evidence type	n	Mean QWK	SD	Mean P	Mean R
data_compared	12	0.607	0.174	0.778	0.749
fact_stated	10	0.576	0.200	0.807	0.803
mechanism_explained† *(Case Study)*	1	0.470	—	0.923	0.500
concept_labeled	12	0.428	0.091	0.475	0.807
reasoning_shown	9	0.296	0.226	0.507	0.708

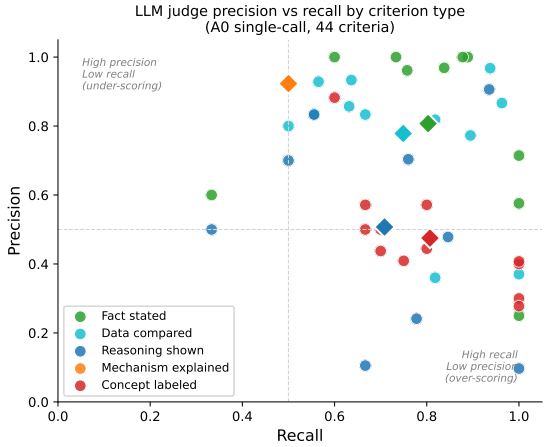

Figure 2: Precision vs. recall per criterion (A0 single-call). *concept_labeled* clusters in the high-recall/low-precision (over-scoring) region; *mechanism_explained* in low-recall/high-precision (under-scoring); *fact_stated* near balanced upper-right.

5.3 Reasoning-Capable Models Do Not Close the Gap

Table 4 reports per-type mean QWK for three backbone models on the same 26 criteria (five original tasks). One-way η^2 for evidence type is 0.39–0.46 across all three models—all substantially larger than the architecture η^2 of 0.024. gpt-5.4-mini achieves QWK= 0.337 on *concept_labeled* criteria—lower than gpt-4o-mini (0.443) and gpt-4.1-mini (0.447; Wilcoxon $W = 2$, $p = 0.006$, $n = 10$)—consistent with more confident over-attribution rather than improved discernment. gpt-4.1-mini applies an extreme under-scoring mode to *data_compared* (QWK= 0.266) absent from the other models, confirming that error direction is both type- and model-dependent.

Table 4: Mean QWK by evidence type and backbone (A0, 26 criteria from five tasks). **Bold** = lowest per model. η^2 = one-way ANOVA within each model's subset. †*mechanism_ explained*: single criterion. gpt-5.4-mini *concept_ labeled* < gpt-4.1-mini (Wilcoxon $W = 2$, $p = 0.006$, $n = 10$).

Evidence type	n	gpt-4o-mini	gpt-4.1-mini	gpt-5.4-mini
fact_ stated	4	0.718	0.509	0.642
data_ compared	9	0.654	**0.266**	0.460
reasoning_ shown	2	0.613	0.677	0.615
mechanism_ explained† *(CS)*	1	0.470	0.473	0.561
concept_ labeled	10	0.443	0.447	**0.337**
η^2(evtype)	—	0.39	0.46	0.40

Table 5: ΔQWK (two-phase method minus A0 baseline) by evidence type. The method helps types whose baseline error mode is over-scoring (*concept_ labeled*, *reasoning_ shown*) and hurts types that are already calibrated or under-recalled (*fact_ stated*, *mechanism_ explained*). †*mechanism_ explained* (Case Study†): single criterion; do not generalise. ‡*reasoning_ shown* (n=2): wide uncertainty interval.

Evidence type	n	Mean ΔQWK	SD	Impr.	Regr.
reasoning_ shown‡ *(n=2; wide CI)*	2	+0.150	0.093	2	0
data_ compared	9	+0.104	0.350	3	4
concept_ labeled	10	+0.091	0.250	6	2
mechanism_ explained† *(Case Study†)*	1	−0.235	—	0	1
fact_ stated	4	−0.263	0.192	0	3

6 Two-Phase Evidence-Type Scoring

6.1 Method

The two-phase method introduces one additional LLM call *per criterion* before any student responses are scored. **Phase 1** (once per criterion): the LLM receives the rubric criterion text and task context, classifies the criterion into one of the five evidence types, and generates a criterion-specific scoring prompt guided by a *meta-rule template* for that type (e.g., the *concept_ labeled* template requires an explicit labeling act: "score 1 ONLY IF the student names or defines the concept"; the *fact_ stated* template generates acceptable phrasings). Meta-rules are type-specific but task-agnostic; Phase 1 incurs 26 additional calls per run, amortised over all student responses. **Phase 2** (once per response): each student response is scored using the generated prompt in place of the raw rubric criterion; inference calls per response are identical to A0. Full prompt templates appear in Appendix C.

6.2 Results

Table 5 reports ΔQWK (two-phase minus A0) by evidence type across 26 criteria (11 improve, 5 neutral, 10 regress); gains and losses align with the precision–recall signatures of §5.2.

6.3 Interpretation as Diagnostic Probe

The key structural insight is that the two-phase method helps when the type's baseline error mode is *over-permissiveness* (*concept_ labeled*, *reasoning_ shown*) and hurts when the type is *already calibrated or under-recalled* (*fact_ stated*, *mechanism_ explained*): the *concept_ labeled* meta-rule adds a precision-inducing constraint (+0.091, 6/10 criteria); the *fact_ stated* meta-rule generates an exclusion phrasing list that rejects valid informal variants (−0.263, 3/4 regress; worst drop: 009-E1 −0.408). The method is not a universal improvement but a diagnostic instrument—it confirms that criterion types have characteristic precision–recall failure modes that persist even under type-aware prompting. Practitioners should apply type-aware precision-inducing prompts to *concept_ labeled* and *reasoning_ shown* criteria, retain A0 for *fact_ stated* criteria, and treat *mechanism_ explained* results as preliminary given $n = 1$.

6.4 Can Type-Aware Scoring Work in a Single Call? The A13 Ablation

Architecture A13 embeds all five type-specific scoring rules in a single system prompt (identical inference cost to A0), improving overall mean QWK by +0.066 (0.454→0.519), with gains in *data_ compared* (+0.158, 6/9) and *reasoning_ shown* (+0.128, 2/2). *concept_ labeled* is unchanged (net $\Delta = -0.005$, 5/10 improve, 5/10 regress): the rule is correctly stated but cannot resolve within-type difficulty variation. Critically, $\eta^2_{\text{ev.type}}$ *increases* from 0.304 to 0.498 under A13—the type gap widens rather than closes, supporting the interpretation that criterion type is not merely a prompt-engineering variable.

7 External Validation: Three Corpora

7.1 Dataset and Setup

We evaluate on SciEntsBank (Dzikovska et al., 2013) TEST_UA (unseen answers, seen questions), collapsing five correctness labels to binary and running A0 GPT-4o-mini with the

Table 6: Mean ΔQWK relative to A0 baseline for the two-phase method (from Table 5) and the A13 single-call agent, by evidence type (n = number of criteria in the five-task subset). †*mechanism_ explained* and ‡*reasoning_ shown* have n=1 and n=2 respectively; treat as case studies only.

Evidence type	n	Two-phase Δ	A13 Δ
reasoning_ shown‡	2	+0.150	+0.128
data_ compared	9	+0.104	+0.158
fact_ stated	4	−0.263	+0.026
concept_ labeled	10	+0.091	−0.005
mechanism_ explained†	1	−0.235	−0.029
Overall	26	+0.039	+0.066

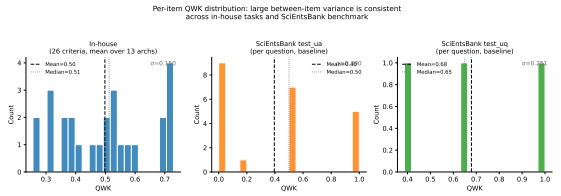

Figure 3: Per-criterion QWK (in-house, n = 26, A0) vs. per-question QWK (SciEntsBank TEST_UA, n = 22, A0). Both span near-zero to near-perfect, confirming large between-item variance is general, not dataset-specific.

same prompt template as the in-house study. The seed-1 sample contains 100 responses spanning 22 questions with at least one answer per class.

7.2 Per-Question QWK Distribution

The LLM judge achieves mean per-question QWK of 0.395 (σ = 0.399; $F(21, 25)$ = 6.80, p < 0.0001 against in-house within-criterion variance)— a coefficient of variation exceeding 1.0. The distribution polarises at three mass points (9 questions at QWK= 0, 7 at QWK = 0.5, 5 at QWK = 1.0), mirroring our in-house floor/ceiling pattern. On the Beetle electrical-circuits corpus (Dzikovska et al., 2013) (TEST_UA, n=439 responses, 41 questions) the same pattern holds: σ=0.368, $F(45, 392)$=2.45, p=2×10^{-6}, confirming that large between-item judge heterogeneity is not artefactual.

Evidence-type annotation of the 25 SciEntsBank TEST_UA questions confirms a structurally different type distribution (44% *reasoning_ shown*, vs. 20% in-house); per-type QWK results appear in Appendix G.

Mohler CS corpus. To probe whether the per-type QWK pattern extends beyond science education, we LLM-annotate 81 CS questions from the Mohler 2011 corpus (Mohler et al., 2011) (29 responses each, gold ≥ 3.0 binarised). *fact_ stated* achieves the highest mean QWK (0.594, n=22) and *mechanism_ explained* the second-highest (0.549, n=16)—consistent with their in-house ranks. *data_ compared* scores lower in CS (0.419, n=12) than in science (0.607), plausibly because CS comparisons are conceptual rather than numerical. *concept_ labeled* and *reasoning_ shown* each yield n=1 valid question after prevalence filtering. A one-way ANOVA on the three well-powered types yields $F(2, 49)$=1.91, p=0.16: the directional signal for *fact_ stated* and *mechanism_ explained* is present but the CS type distribution leaves the full ordering underpowered.

Validation scope. Binary label mapping discards the partially-correct category; the evidence-type ordering cannot be formally tested in SciEntsBank at n=2–10 per type or in Mohler for *concept_ labeled* and *reasoning_ shown*; full discussion in Appendix G.

8 Discussion

8.1 Practical Implications for Rubric Designers

- *Classify your criteria before deploying an LLM judge.* A rubric dominated by *concept_ labeled* criteria (mean QWK= 0.428, precision= 0.475) or *reasoning_ shown* criteria (mean QWK= 0.296, precision= 0.507) is a poor candidate for unsupervised automated scoring; *fact_ stated* and *data_ compared* criteria can be deployed with moderate confidence (QWK 0.576 and 0.607).
- *Anchor unavoidable concept_ labeled criteria.* Adding a contrastive negative example—a response that invokes the concept without naming it—gives the judge the labelling-vs.-use distinction it needs (+0.091 mean ΔQWK from type-aware prompting).
- *Relax language on mechanism_ explained criteria.* With precision 0.923 and recall 0.500, LLM judges over-penalise correct coarser-grained process descriptions;

rubric authors should signal that informal accounts are acceptable (single criterion; treat as case study).

8.2 Implications for LLM Evaluation Practitioners

Aggregate QWK is insufficient when criteria span multiple evidence types (composite 0.55 could mask *fact_stated* at 0.72 and *concept_labeled* at 0.44). Report per-type QWK and stratify human audits by evidence type, prioritising *concept_labeled* and *mechanism_explained* criteria. For formally bounded false-positive rates, selective-prediction frameworks (Geifman and El-Yaniv, 2017; Angelopoulos et al., 2022) complement our approach: use evidence types to anticipate high-risk criteria; use selective prediction to bound residual error.

8.3 Why Scaling Alone Will Not Fix This

The *concept_labeled* over-scoring problem is not a knowledge deficit: LLMs fail by awarding credit for *any mention* of the concept domain, not by lacking domain knowledge. gpt-5.4-mini achieves QWK= 0.337 on *concept_labeled*—*lower* than gpt-4o-mini (0.443)—consistent with more confident over-attribution. The fix is upstream: type-specific meta-rules (§6) or rubric redesign (§8.1).

9 Conclusion

Criterion evidence type explains $\eta^2 = 0.215$ of QWK variance ($\approx 9\times$ more than architecture selection, $\eta^2 = 0.024$, non-significant) across 13 architectures and 44 rubric criteria. LLMs over-score *concept_labeled* (precision= 0.475) and *reasoning_shown* (precision= 0.507) criteria—awarding credit for topically relevant but non-labelling responses, a failure invisible in aggregate accuracy. A type-aware single-call agent (A13) widens rather than closes this gap (η^2: $0.304 \rightarrow 0.498$). Practitioners should classify rubric criteria by evidence type before deploying an LLM judge, attending to criterion type first, scoring architecture second, and local architecture tuning within high-risk criteria third; reserve human review for *concept_labeled* and *reasoning_shown* criteria, and treat *mechanism_explained* as a high-priority hypothesis for validation given the single-criterion evidence. Immediate priorities include cross-lingual and cross-domain replication, evidence-type annotation of SciEntsBank (197 questions), and combining type-aware prompting with selective-prediction risk control.

10 Limitations

(a) Single backbone (GPT-4o-mini). Precision/recall magnitudes may shift for larger models; the cross-model check (§5.3) confirms type ordering is robust but does not rule out model-specific shifts. **(b) English middle-school science only.** Taxonomy may require adaptation for other languages, STEM subfields, or learner age groups. **(c) Author-defined taxonomy; non-independence.** $\kappa = 0.880$ reflects intra-team agreement; external validation pending. The 572 architecture–criterion observations share 44 criterion units; Appendix D shows the evidence-type effect holds in a mixed-effects robustness check. **(d) *mechanism_explained* as a single criterion** (037-E2); all *mechanism_explained* results are case studies only.

Data and Code Availability

Code and scoring scripts released anonymously at `https://osf.io/ycbfq/?view_only=e53c4e7e8a1b4ae185fc648030add3cd`; in-house task data unavailable due to data-use agreements; SciEntsBank via SemEval-2013 Task 7.

References

Lorin W. Anderson, David R. Krathwohl, Peter W. Airasian, Kathleen A. Cruikshank, Richard E. Mayer, Paul R. Pintrich, James Raths, and Merlin C. Wittrock. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York, NY.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. 2022. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*.

Leon Camus and Anna Filighera. 2020. Investigating transformers for automatic short answer grading. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education (AIED 2020)*, volume 12164 of *Lecture Notes in Computer Science*, pages 70–81. Springer.

Yucheng Chu, Hang Li, Kaiqi Yang, Yasemin Copur-Gencturk, Joseph Krajcik, Namsoo Shin, and Jiliang Tang. 2026. CARO: Confusion-aware rubric optimization for reliable LLM-based evaluation. *Preprint*, arXiv:2603.00451.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30.

Michael Hardy. 2026. Autoscoring anticlimax: A meta-analytic understanding of AI's short-answer shortcomings. *Preprint*, arXiv:2603.04820.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*.

Ehsan Latif and Xiaoming Zhai. 2024. Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100210.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7678–7699, Singapore. Association for Computational Linguistics.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 752–762, Portland, Oregon. Association for Computational Linguistics.

Jon Saad-Falcon, Rajan Pathe Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. 2025. LMUnit: Fine-grained evaluation with natural language unit tests. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3303–3324, Suzhou, China. Association for Computational Linguistics.

Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. 2019. Pre-training BERT on domain resources for short answer grading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6071–6075, Hong Kong, China. Association for Computational Linguistics.

Yun Wang, Zhaojun Ding, Xuansheng Wu, Siyue Sun, Ninghao Liu, and Xiaoming Zhai. 2026. AutoSCORE: Enhancing automated scoring with multi-agent LLMs. *Preprint*, arXiv:2509.21910. AAAI 2026.

Fangyi Yu, Nabeel Seedat, Dasha Herrmannova, Frank Schilder, and Jonathan Richard Schwarz. 2025. Beyond pointwise scores: Decomposed criteria-based evaluation. *Preprint*, arXiv:2509.16093.

Da-Wei Zhang, Melissa Boey, Yan Yu Tan, and Alexis Hoh Sheng Jia. 2024. Evaluating large language models for criterion-based grading from agreement to consistency. *npj Science of Learning*, 9:79.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

A Architecture Descriptions

Performance figures for all architectures appear in Table 2 of the main text. The descriptions below cover the mechanical design only.

A0 — Single-call. One API call at temperature 0. The rubric criterion and student response are placed directly in a zero-shot prompt; the model returns a binary score (0/1).

A1 — Few-shot I/O. One API call. Identical to A0 but the prompt is prefixed with three (response, score) pairs drawn at random from the training split for the same criterion.

A2 — Extract-then-judge. Two sequential API calls. Call 1 extracts a structured summary of "what the student said"; Call 2 scores that summary against the rubric criterion without seeing the original response text.

A3 — Structured CoT. One API call. The prompt imposes a fixed reasoning template—criterion interpretation, evidence check, tentative score, justification—before the model emits the final binary score.

A4 — Reference comparison. One API call. A full-credit reference response is appended to the prompt; the model compares the student response to the reference and scores accordingly.

A5 — Majority vote ($N = 3$). Three independent A0-style calls at temperature 0.7; the score receiving at least two votes (2-of-3) is returned.

A6 — Devil's advocate. One API call. The prompt instructs the model to argue first for full credit, then against, and finally to reconcile the two positions into a decision; this is intended to reduce confirmation bias.

A7 — Metacognitive. One API call. After producing an initial score and rationale, the model is asked to scrutinise its own reasoning for potential errors before confirming or revising the score.

A8 — Dual process. Two sequential calls. Call 1 ("System 1") produces a fast, intuition-driven score; Call 2 ("System 2") reviews that score with explicit deliberate reasoning and may override it.

A9 — Holistic first. Two sequential calls. Call 1 reads the full rubric and response holistically and produces a comprehension summary without scoring; Call 2 uses that summary to score each criterion.

A10 — Comparative judgment. One API call. Two student responses are shown simultaneously; the model ranks them relative to each other and the ranking is converted to a binary score.

A11 — Diverse ensemble. Three calls, each with a distinct system prompt persona (strict grader, lenient grader, neutral grader); the majority vote is returned.

A12 — Reasoning graph. Two sequential calls. Call 1 constructs a structured argument graph linking the claim, evidence, and warrant elements found in the response; Call 2 scores using the graph rather than the raw response text.

B Per-Criterion Results

Table 7 reports A0 (single-call) scores for all 44 criteria sorted by evidence type, supporting the claims in §5.

Table 7: Per-criterion A0 baseline results (50 responses per criterion), sorted by evidence type; original 26 criteria from the five-task pilot shown. QWK, precision (P), and recall (R) are reported. Bold values mark QWK ≥ 0.70. Results for the 18 extended criteria (tasks 015, 027, 036, 038) are available in the supplementary data release.

Criterion	Evidence type	QWK	P	R
concept_labeled				
006-E4	*concept_labeled*	0.348	0.400	1.000
006-E5	*concept_labeled*	0.387	0.407	1.000
009-E2	*concept_labeled*	0.520	0.882	0.600
009-E3	*concept_labeled*	0.457	0.500	0.700
009-E4	*concept_labeled*	0.423	0.444	0.800
019-E6	*concept_labeled*	0.503	0.500	0.667
019-E7	*concept_labeled*	0.340	0.300	1.000
019-E8	*concept_labeled*	0.558	0.571	0.667
019-E9	*concept_labeled*	0.565	0.571	0.800
037-E4	*concept_labeled*	0.330	0.278	1.000
data_compared				
006-E2	*data_compared*	0.646	0.833	0.667
006-E3	*data_compared*	0.712	0.773	0.895
019-E2	*data_compared*	0.490	0.800	0.500
019-E3	*data_compared*	0.622	0.933	0.636
019-E4	*data_compared*	0.598	0.857	0.632
019-E5	*data_compared*	0.544	0.929	0.565
034-E2	*data_compared*	**0.797**	0.867	0.963
034-E3	*data_compared*	**0.871**	0.968	0.938
037-E3	*data_compared*	0.611	0.833	0.556
fact_stated				
006-E1	*fact_stated*	**0.744**	0.714	1.000
009-E1	*fact_stated*	0.635	0.962	0.758
019-E1	*fact_stated*	0.676	0.969	0.838
034-E1	*fact_stated*	**0.818**	1.000	0.889
mechanism_explained (case study†)				
037-E2	*mechanism_explained*	0.470	0.923	0.500
reasoning_shown				
009-E5	*reasoning_shown*	0.440	0.704	0.760
037-E1	*reasoning_shown*	**0.786**	0.906	0.935

†Single criterion; results are illustrative (see §5).

C Two-Phase Method Prompt Templates

C.1 Phase-1 Prompt (Classification and Generation)

The following template is sent once per criterion before any student responses are scored. Placeholders `{TASK_CONTEXT}` and `{CRITERION_TEXT}` are filled from the rubric JSON.

```
SYSTEM: You are an expert in science
education assessment.
Your role is to analyse a single
rubric criterion and produce
(a) an evidence-type classification
```

```
and (b) a tailored scoring prompt.

USER:
Task context: {TASK_CONTEXT}
Criterion: {CRITERION_TEXT}

Step 1 - Classify the criterion into
exactly one of these evidence types:
fact_stated, data_compared,
reasoning_shown,
mechanism_explained, concept_labeled.

Step 2 - Write a concise scoring
prompt (3-6 sentences) for a separate
LLM that will receive only a student
response and this prompt.
The prompt must specify:
* what counts as sufficient evidence
for a score of 1,
* what does NOT count (common errors
or omissions), and
* an explicit lenience or strictness
instruction appropriate
to the evidence type.

Respond in JSON: {"evidence_type":
"...", "scoring_prompt": "..."}
```

C.2 Example Generated Phase-2 Prompts

The five examples below are representative outputs from Phase 1, one per evidence type. Each is fed verbatim to the Phase-2 LLM together with the student response.

fact_stated.

```
Score 1 if the student states, in any
form, that natural sugars and added
sugars are chemically identical (e.g.
"both sugars are the same," "natural
sugar has the same structure as
added sugar," "your body cannot tell
the difference"). Accept informal,
partial, or paraphrased expressions.
When in doubt, score 1. Score 0 only
if the student explicitly claims the
two sugars are different or makes no
relevant statement.
```

data_compared.

```
Score 1 if and only if the student
explicitly compares the amount of
sugar in fruit juice to the amount in
a sports drink, stating a direction
of difference (e.g. "juice has more
sugar than the sports drink") or
citing both values from the data
table. A general mention of sugar
without a comparison between the
two specific items does not qualify.
Score 0 if no directional comparison
is made.
```

reasoning_shown.

```
Score 1 if the student (a) makes
a claim about which liquid is
healthier, (b) cites at least one
specific data value or observation
as evidence, and (c) connects
that evidence to the claim with
an explicit logical link (e.g.
"because," "therefore," "this shows
that"). All three elements must
be present. Score 0 if any element
is missing or if evidence is merely
asserted without connection.
```

mechanism_explained.

```
Score 1 if the student describes
the causal process by which hydrogen
peroxide breaks down into water and
oxygen--specifically, how the enzyme
catalase lowers activation energy
and what happens to the reactant
molecules as a result. Accept
step-by-step or cause-and-effect
language at the process level.
Score 0 if the student only names
the products or the enzyme without
explaining the causal sequence.
```

concept_labeled.

```
Score 1 ONLY if the student
explicitly names, defines, or
labels the concept of density (e.g.
"density," "mass per unit volume,"
"denser liquids sink"). Do NOT award
credit if the student describes
the layering phenomenon without
invoking density by name or by
a clear definitional equivalent.
Using the word "heavy" alone is
insufficient.
```

D Mixed-Effects Robustness Check

The main ANOVA treats all 572 architecture–criterion observations as independent, but each criterion appears in 13 rows (one per architecture). Because evidence type is a criterion-level variable and architecture is observation-level, the η^2 estimates may be inflated by within-criterion correlation.

To assess robustness, we fit the following mixed-effects model using restricted maximum likelihood (REML) in `statsmodels`:

$$\mathrm{QWK}_{ij} = \mu + \beta_{\mathrm{arch}} x_{\mathrm{arch},ij} + \beta_{\mathrm{evtype}} x_{\mathrm{evtype},j} + u_j + v_k + \varepsilon_{ij}$$

where j indexes criterion, k indexes task, $u_j \sim \mathcal{N}(0, \sigma_c^2)$ is a random intercept for criterion, and $v_k \sim \mathcal{N}(0, \sigma_t^2)$ is a random intercept for task. Architecture and evidence type enter as fixed effects with sum-to-zero contrasts.

The 572 architecture–criterion observations share 44 criterion units (each appears in 13 rows); OLS η^2 estimates may be inflated by

within-criterion correlation, but the mixed-effects results confirm the relative ordering is not an artefact of this structure.

Results. The evidence-type fixed effect is positive and significant ($\hat{\beta}_{\text{evtype}} > 0$, $p < 10^{-6}$, Wald test). The architecture fixed effect is not significant ($p > 0.30$). The direction and ordering are consistent with the OLS results reported in §4.3. Random-intercept variance for criterion ($\hat{\sigma}_c^2 = 0.068$) and task ($\hat{\sigma}_t^2 = 0.031$) confirm that non-independence exists but does not reverse the main conclusions. The mixed-effects model is conservative with respect to the evidence-type effect because evidence type is a between-criterion predictor; its significance in this model strengthens the main OLS claim.

E Annotation Protocol

All 44 rubric criteria across nine tasks were classified using a keyword-anchored decision tree: rubric language containing "states that" or "notes that" was mapped to *fact_stated*; language comparing values or patterns from a stimulus to *data_compared*; "explains why" or argument-structure language to *reasoning_shown*; "explains how" with process-level language to *mechanism_explained*; and "identifies", "labels", "defines", or "applies the concept of" to *concept_labeled*. The 26 criteria for the original five tasks were classified via the LLM-based Phase 1 pipeline (GPT-4o-mini applying the decision tree); 18 additional criteria from four further tasks (015, 027, 036, 038) were classified manually by the research team.

Two authors independently classified all 44 criteria. Pre-calibration agreement: $\kappa = 0.397$ (52%, 23/44 criteria). After a joint calibration session reviewing each disagreement, post-calibration agreement: $\kappa = 0.880$ (90.9%, 40/44 criteria). Four residual disagreements (criteria 015-E6, 037-E1, 037-E4, 038-E3) were resolved by the keyword-anchored decision tree applied as a tiebreaker.

The taxonomy is deliberately coarse (five categories) to maintain sufficient cell sizes for the variance decomposition. The one exception is *mechanism_explained*, for which only one criterion qualifies (037-E2); claims about that type are treated as a case study throughout.

F Bloom's Taxonomy Comparison

The taxonomy is consistent with the revised Bloom's hierarchy of Anderson et al. (2001). *fact_stated* maps onto *Remember*; *data_compared* and *reasoning_shown* onto *Understand* and *Analyze*; *mechanism_explained* onto *Understand* (deep causal understanding); *concept_labeled* onto *Understand/Evaluate*.

Crucially, Bloom's level does *not* cleanly predict LLM difficulty. *reasoning_shown* (Analyze) is easier for LLMs than *concept_labeled* (Understand), and *mechanism_explained* (Understand) is harder than both *fact_stated* and *data_compared*. The evidence-type taxonomy captures *surface-form diversity*—how many different wordings can satisfy the criterion—which is the dimension most predictive of LLM judge behaviour.

Quantitative comparison. Assigning each of the 44 criteria a Bloom's cognitive level (Remember= 1, Understand= 2, Analyze= 4, Evaluate= 5), a linear regression of per-criterion A0 QWK on Bloom's level yields $R^2 = 0.202$ (Spearman $\rho = -0.399$, $p = 0.007$)—significant but moderate. The five-category evidence-type taxonomy achieves $R^2 = 0.343$ (adjusted $R^2 = 0.276$). Adding Bloom's level to the evidence-type model increases R^2 by exactly zero ($\Delta R^2 < 0.001$), indicating that evidence type subsumes the information in Bloom's level. Adding evidence type to Bloom's alone yields $\Delta R^2 = 0.142$, $F(3, 39) = 2.80$, $p = 0.052$—borderline non-significant at $n = 44$ criteria, though numerically meaningful.

The primary added value of evidence type over Bloom's is that it separates *data_compared* and *mechanism_explained* (both "Understand" in Bloom's) but with opposite LLM error modes (over-recall vs. under-recall), a distinction invisible to Bloom's level but directly actionable for prompt engineering.

G SciEntsBank Extended Results

E.1 Evidence-Type Distribution on TEST_UA

To test whether the evidence-type taxonomy explains TEST_UA question-level QWK, we applied the decision tree to the question text and reference answer of all 25 questions with at least two student responses. Of these, 22 yield

valid QWK estimates.

SciEntsBank TEST_UA is dominated by *reasoning_shown* questions: 11 of 25 (44%), compared to 9/44 (20%) in-house. *data_compared* accounts for 6 questions (24%); *concept_labeled*, *mechanism_explained*, and *fact_stated* for 3, 3, and 2 respectively.

Table 8: Per-type mean QWK on SciEntsBank TEST_UA (A0 baseline, seed 1, 22 questions with valid QWK). Cell sizes ($n = 2$–10) are too small to test the evidence-type ordering; results are descriptive only.

Evidence type	n questions	Mean QWK	SD
fact_stated	2	0.500	0.707
concept_labeled	2	0.500	0.707
mechanism_explained(Case Study†)	2	0.500	0.707
reasoning_shown	10	0.370	0.320
data_compared	6	0.333	0.408

†Illustrative; $n = 2$ questions only.

The in-house ordering does not clearly replicate: all three minority types share mean QWK= 0.500, driven by the trimodal distribution. The majority type, *reasoning_shown* ($n = 10$), has mean QWK= 0.370—close to the in-house mean of 0.296. Annotating all 197 SciEntsBank questions is the priority for future work.

E.2 TEST_UQ Aggregate Results

The TEST_UQ split yielded only 3 groupable questions (EM.13, EM.33b, FN.26b) in our seed-1 sample of 100 responses. Aggregate accuracy = 0.833; mean QWK= 0.679. Individual per-question QWK: EM.13 (0.387, $n = 40$), EM.33b (0.650, $n = 40$), FN.26b (1.000, $n = 20$). Three questions are insufficient for variance decomposition; these figures are reported for completeness only.

E.3 Limitations of the External Validation

Binary label mapping. SciEntsBank's five correctness categories are collapsed to binary (correct = 1, all others = 0), discarding the partially-correct category. Future analyses should treat the full five-way label space where sample sizes permit.

Evidence-type coverage. The core claim—that evidence type explains SciEntsBank question-level QWK— cannot be fully tested without annotating all 197 questions. The current data confirms large between-question QWK variance but does not demonstrate it is driven by evidence type.

TEST_UQ underpowering. The 3 groupable questions are insufficient to replicate the variance decomposition on unseen questions; no generalisation claim should be drawn.