Evaluating Logical Error Sensitivity in Text-Only Proxies for Generated Science Rubrics

Lauren Mercer, Maya Kalyani Reed, Hoang Justin Tran, and Alexander J. Garnier

[1] Simon Fraser University, lauren.mercer@sfu.ca
[2] University of Missouri, Columbia, maya.kalyani.reed@umsystem.edu
[3] Université Grenoble Alpes, hoang.justin.tran@etu.univ-grenoble-alpes.fr
[4] Concordia University, alexander.j.garnier@concordia.ca

Abstract. Automatically generated marking rubrics could reduce assessment-authoring effort, but educational usefulness depends on whether evaluation detects criteria that would misgrade student responses, not only fluent paraphrases. This study examines that measurement problem in a public, text-only proxy testbed built from SciQ support explanations; the references are constructed proxies, not educator-authored rubrics. A locally served Qwen3-14B instruction model generated two-point rubrics for 120 held-out science questions under three decoding seeds and was compared with answer-only, retrieval-edit, and reference-template controls. Its outputs contained the supplied answer on 0.992 of items, while support-concept coverage was 0.263 to 0.268 and point-F1 was 0.434 to 0.436. An answer-only baseline reached point-F1 0.547 under the same constructed metric. Deliberately negated rubrics still had lexical cosine 0.901 and ROUGE-L 0.849; the concept metric remained 0.724 because polarity was not represented. Across 360 model generations, reference-similarity scores correlated with point-F1. These corruption tests are not calibrated acceptance decisions, but they show that aggregate association is insufficient evidence of logical sensitivity. The study provides code, data hashes, raw generations, repeated-seed estimates, and controlled corruptions for developing evaluation protocols that test negation, omission, irrelevant support, and unsupported generic language.

Keywords: AI education, assessment, scholarly integrity

1 Introduction

Science assessments often rely on rubrics that identify credit-bearing concepts, and generative systems could lower authoring costs only if their outputs preserve the logic needed for consistent grading. Standard text-similarity measures are attractive for scalable evaluation, yet they may reward surface plausibility even when a criterion omits a necessary idea, reverses polarity, or introduces irrelevant support. This work contributes a bounded, reproducible learning-science case study of that evaluation problem. Using a text-only proxy constructed from

science-question support explanations rather than educator-authored rubrics, it compares one open instruction model with deterministic controls over repeated seeds and uses targeted corruptions to separate average reference similarity from sensitivity to consequential rubric errors. Its evidence supports evaluation protocols that include explicit checks for negation, omission, unsupported generic language, and unrelated support.

Structured science questions are often scored with marking rubrics that divide a response into credit-bearing concepts. Producing those criteria is labor-intensive, and automatically generated rubrics could assist instructors or assessment designers. The central difficulty is evaluation. A generated criterion can be fluent and close to a reference while omitting a necessary concept, reversing its polarity, or adding an irrelevant mechanism. These are not merely stylistic defects: each can change which student response receives credit.

Wei et al. recently proposed a multimodal sequence-to-sequence system that combines a Bidirectional Encoder Representations from Transformers (BERT) text encoder, a ResNet50 diagram encoder, and a gated recurrent unit decoder for science marking-scheme generation [4]. Their reported contrast between low BLEU and near-unity BERT-based similarity motivates an important field question: which automatic measures can establish that a generated rubric is faithful enough to support grading? Legitimate paraphrases may have low word overlap, yet high reference similarity can also coexist with failures that teachers would reject. Resolving this distinction requires baselines and targeted perturbations in addition to an average similarity score.

The present study isolates this measurement problem in an open setting. SciQ provides science questions, answers, distractors, and supporting explanations [40]. We convert each support explanation into a two-part reference that asks for the correct answer and its supporting scientific idea. This construction is not an educator-authored rubric and is not presented as one. It provides an inspectable proxy testbed in which the references, generation procedure, and failure transformations can be rerun. We compare Qwen3-14B with deterministic controls, measure variation across three seeds, and introduce omission, negation, generic-language, and unrelated-support corruptions.

Automated short-answer grading and large-language-model (LLM) assessment studies increasingly report embedding or generation metrics alongside task scores [5, 14, 25, 39, 48]. This development makes construct validity especially important: an evaluation measure should respond to errors that matter for the intended educational use. The closest methodological precedent is not a new rubric generator, but behavioral and perturbation-based testing of NLP systems and NLG metrics [24, 30, 34, 45]. Accordingly, the contribution here is a bounded educational case study showing how such failure-oriented evaluation appears in a science-rubric proxy, not a claim to introduce the general stress-testing paradigm.

The paper makes four contributions:

– A public, fully specified proxy testbed for investigating automatic evaluation of generated science rubrics, with data hashes and raw outputs.

- A comparison of one open instruction model with answer-only, retrieval-edit, and reference-template controls over repeated decoding seeds.
- Controlled corruptions that test whether reference-based measures recognize omission, negation, unsupported generic language, and unrelated scientific support.
- An analysis separating average metric correlation from sensitivity to consequential rubric errors.

2 Related work

2.1 Automated assessment of constructed responses

Automatic short-answer grading has been formulated as semantic matching, classification, or regression from a question, a reference answer, and a learner response to a score. Reviews trace a progression from handcrafted similarity features to neural encoders and contextual embeddings [5, 14, 29]. BERT- and XLNet-based approaches, co-attention architectures, and optimized deep models have been evaluated across several languages and educational domains [1, 2, 10, 13]. Biology and other open-ended assessment settings further show that explanations and scoring criteria must be aligned with the content being assessed [3, 32, 42].

Generative assessment systems change the evaluation target. Rather than predicting an existing score, they may produce feedback, questions, or grading criteria. Recent work has examined knowledge-grounded answer evaluation, LLM fine-tuning for short-answer grading, science grading with causal and multimodal analysis, and automated test generation and marking [9, 16, 25, 35, 39, 48]. These systems make fluency easy to observe, but fidelity is harder to establish. A generated marking criterion must be scientifically correct, relevant to the item, sufficiently complete, and operational enough to support consistent credit decisions.

Simple controls remain valuable in this setting. High-performing shallow short-answer systems demonstrate that architectural complexity does not remove the need for direct task comparisons [38]. SemEval resources similarly encourage benchmark-oriented evaluation rather than reliance on one text-similarity statistic [23, 31]. Rubric-based essay grading and assignment-generation work also treats scoring criteria and assessment objectives as the operative units, rather than unrestricted sentence resemblance [18, 22, 36].

2.2 Metrics and stress tests for generated educational text

BLEU measures n-gram precision and was developed for machine translation [26]. In rubric generation, a single reference can penalize valid paraphrases, which motivates embedding-based alternatives. BERTScore and related measures reward contextual semantic correspondence rather than only exact overlap [41, 46]. Wei et al. used the same motivation when pairing BLEU with a BERT-derived similarity measure [4].

Semantic similarity remains a measurement instrument rather than direct evidence of grading validity. Studies of text-generation metrics show that average scores can have limited sensitivity to critical or localized errors [17, 19, 21, 33]. More directly, CheckList-style work evaluates systems or metrics with targeted perturbations rather than aggregate accuracy or correlation alone [24, 30, 34, 45]. Those studies establish the general method used here: controlled changes can reveal whether a metric rewards the intended behavior. The present paper applies that logic to constructed science-rubric references and reports the resulting limitations in an educational assessment setting.

Rubrics create a particularly demanding case because a small linguistic change can invert a scoring rule. For example, inserting "not" can preserve nearly every content word while changing the credit decision. Recent rubric-generation evaluation work also emphasizes expert standards and response-level interventions: RubricBench compares model-generated rubrics with human standards, and Interventional Transfer evaluates whether generated and expert rubrics move together when responses are perturbed to pass or fail criteria [37, 47]. The SciQ proxy here is much smaller and lacks expert-authored atomic rubrics or response interventions, so its role is diagnostic rather than benchmark-level validation.

2.3 Multimodal grounding and prompt-based adaptation

Science assessment can depend on diagrams, graphs, chemical structures, and experimental layouts. Multimodal sequence modeling provides plausible mechanisms for combining visual and textual evidence [6, 27, 28]. The architecture studied by Wei et al. follows this direction with BERT, ResNet50, and a recurrent decoder [4, 8, 15]. Vision-language generation and referring-expression comprehension also show how language can anchor visual distinctions. Evaluation of a diagram-aware grader, however, must identify items that genuinely require the diagram and compare text-only, visual-only, and multimodal conditions.

Prompted generation raises a related measurement question: what supervision is available, and which behavior can an evaluation legitimately attribute to it? Multimodal sequence modeling and BERT/ResNet representation learning show that grounding claims depend on how textual and visual evidence is coupled [8, 15, 27, 28]. Language-anchored visual generation and referring-expression grounding use labels or text as supervisory anchors [11, 20]; model reprogramming and in-context semi-supervised learning examine transfer under constrained supervision [7, 12]; and generative tissue reconstruction evaluates generated structure against domain evidence [44]. This supervision-centered perspective directly informs our proxy: because Qwen3 receives the answer but not the SciQ support passage, answer-hit is interpreted as prompt adherence, whereas support coverage measures reproduction of withheld reference evidence. The resulting distinction among available supervision, transferred structure, and evaluation evidence organizes the controls used here.

In the present experiment, the fixed prompt and retrieval-edit control serve as modest prompt-based and retrieval-based comparators. They are not substitutes

for a fine-tuned educational model, but they help reveal what the instruction model adds beyond concise answer reproduction or reuse of a related training item.

Wei et al. supply the motivating problem and the observation that lexical and embedding measures can disagree [4]. This study takes that evaluation question into a public text-only setting. Its new elements are deterministic controls, repeated seeds, exact metric definitions, raw generations, and controlled error transformations. It neither reuses the original data nor claims to test its visual encoder.

3 Methods

3.1 Study scope and data

We study whether commonly reported reference-based measures respond to consequential errors in generated rubric-like text under a public, text-only SciQ protocol. The scope is evaluation design; it does not include classroom deployment, diagram comprehension, or claims that the constructed references are teacher-authored marking schemes.

SciQ contains multiple-choice science questions, four answer options, a correct answer, and a support explanation [40]. The distribution used here contains 11679 training examples, 1000 validation examples, and 1000 test examples. The three public parquet files were downloaded, and their byte sizes and SHA256 hashes were recorded. The evaluation uses the first 120 test items. This subset was chosen as a small rerunnable proxy rather than by random or stratified sampling.

For each item, the reference has two criteria: identify the correct answer and state the supporting scientific idea from the SciQ support field. This deterministic transformation makes the target inspectable. It also creates a known bias: lexical concept measures favor language that reproduces the support passage and may under-credit a scientifically correct paraphrase. Fifteen of the selected items have empty support fields, so their constructed references contain only the answer criterion. The template control returns this constructed reference and is therefore an upper anchor for the implemented metrics, not a deployable system.

3.2 Systems

The answer-only control states that the correct answer has been identified. It tests whether a longer output contributes reference concepts beyond answer reproduction. The retrieval-edit control identifies the most similar training question using term frequency–inverse document frequency over the question stem, takes its constructed reference, and replaces the answer string. It tests whether reusing assessment language from a nearby item is competitive with direct generation. These controls are diagnostics for the metric, not realistic educational baselines.

Qwen3-14B receives a fixed instruction to produce two semicolon-separated marking points for each question. Qwen3 is an open model family documented in a technical report [43]. All calls used the same local OpenAI-compatible service with model name `qwen3-14b`, temperature 0.3, top-p 0.9, maximum output 256 tokens, and seeds 0, 1, and 2. The run contained 360 model calls, and the raw request-response records are retained as JSONL. The endpoint address is recorded in the accompanying provenance file.

3.3 Metrics

Answer-hit equals one when the answer tokens occur in the prediction. Because the prompt supplies the correct answer, answer-hit describes prompt adherence rather than answer derivation. Support coverage is the fraction of non-stopword support tokens found in the prediction. Point precision is the fraction of prediction content tokens contained in the constructed reference concept set. Point recall averages answer-hit and support coverage. Point-F1 is the harmonic mean of point precision and point recall. This definition gives answer and support equal weight regardless of how many concepts appear in the support, penalizes valid vocabulary outside the constructed reference, and does not encode polarity. It can also miss paraphrases: for example, a model output that says "a neutral solution" does not trigger answer-hit for the SciQ answer string "a neutral one." These transparent measures evaluate reproduction of the constructed answer and support concepts; they are implementation-specific diagnostics rather than validated measures of rubric quality or expert judgment.

We additionally report sentence-level BLEU, ROUGE-L F1, and lexical cosine over bag-of-token counts. Lexical cosine is not BERTScore. It is used as a fully specified similarity measure because the BERT-similarity implementation reported by Wei et al. is not specified in sufficient detail to reconstruct [4]. All intervals in the system and corruption tables are item-level 95% Student intervals.

3.4 Controlled corruptions

Four transformations probe distinct failure modes. Omission retains only the answer criterion. Negation states that the answer should not be identified and that the supporting idea is false. Generic language retains the answer but replaces the explanation with a non-specific phrase. Unrelated support retains the answer and appends an irrelevant mechanism. These transformations are defined to be invalid relative to the constructed reference, so a measure intended to support rubric validation should decrease when they are applied. They are deliberately simple string transformations. They are not sampled from model errors, and the unrelated-support phrase is a placeholder rather than a plausible misconception. On the 15 items with empty support, omission is identical to the constructed reference.

Table 1. System comparison. Values are means ± 95% confidence intervals over 120 items.

System	Answer-hit	Support coverage	Point-F1	BLEU	ROUGE-L	Lexical cosine
Template reference	1.000 ± 0.000	0.875 ± 0.060	0.814 ± 0.027	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Answer-only	1.000 ± 0.000	0.071 ± 0.013	0.547 ± 0.009	0.094 ± 0.041	0.384 ± 0.039	0.601 ± 0.026
Retrieval-edit	1.000 ± 0.000	0.220 ± 0.041	0.320 ± 0.031	0.348 ± 0.039	0.503 ± 0.033	0.672 ± 0.028
Qwen3-14B, seed 0	0.992 ± 0.017	0.263 ± 0.039	0.434 ± 0.035	0.049 ± 0.009	0.253 ± 0.019	0.475 ± 0.026
Qwen3-14B, seed 1	0.992 ± 0.017	0.268 ± 0.040	0.436 ± 0.035	0.050 ± 0.009	0.254 ± 0.019	0.477 ± 0.026
Qwen3-14B, seed 2	0.992 ± 0.017	0.265 ± 0.040	0.435 ± 0.035	0.049 ± 0.008	0.254 ± 0.019	0.476 ± 0.026

3.5 Analysis and reproducibility

We report system means by seed and correlate each reference-based measure with point-F1 over the model generations. The correlation analysis tests average association, whereas the controlled corruptions test local failure sensitivity. The SciQ split files, model outputs, result tables, figure-generation procedure, and endpoint metadata are preserved with the analysis package. Sixty-nine of the 120 items produced identical Qwen3-14B outputs under all three seeds, and the pooled correlations treat the three outputs per item as descriptive observations rather than independent samples. The full call time recorded by the endpoint was 146.137 seconds, excluding data preparation and metric computation.

4 Results

4.1 Model outputs and deterministic controls

Table 1 reports the system comparison. Qwen3-14B output contained the supplied answer on 0.992 of items for every seed. This near-ceiling value is expected because the correct answer appears in the prompt and does not measure answer derivation. Support coverage ranged from 0.263 to 0.268 and point-F1 from 0.434 to 0.436. The narrow ranges show aggregate stability across the tested decoding seeds, but stability does not establish correctness.

The answer-only control reached point-F1 0.547, above the model rows. That comparison reflects the metric and reference construction as well as the systems. A six-word answer statement is precise under a token-based concept measure, whereas a roughly 23-word generated explanation introduces scientifically plausible words that may not occur in the SciQ support passage. The result therefore does not show that terse answers are better rubrics. It shows that model fluency does not guarantee improvement on the chosen concept-reproduction target and that metric interpretation requires simple controls. The template reference itself has point-F1 0.814 rather than 1.000, further showing that point-F1 is not a normalized rubric-quality scale. The 15 empty-support items make this especially visible: their self-reference point-F1 averages 0.433 because the implementation still averages answer-hit with zero support coverage.

The retrieval-edit control achieved the highest BLEU, ROUGE-L, and lexical cosine among the non-template systems while obtaining the lowest point-F1.

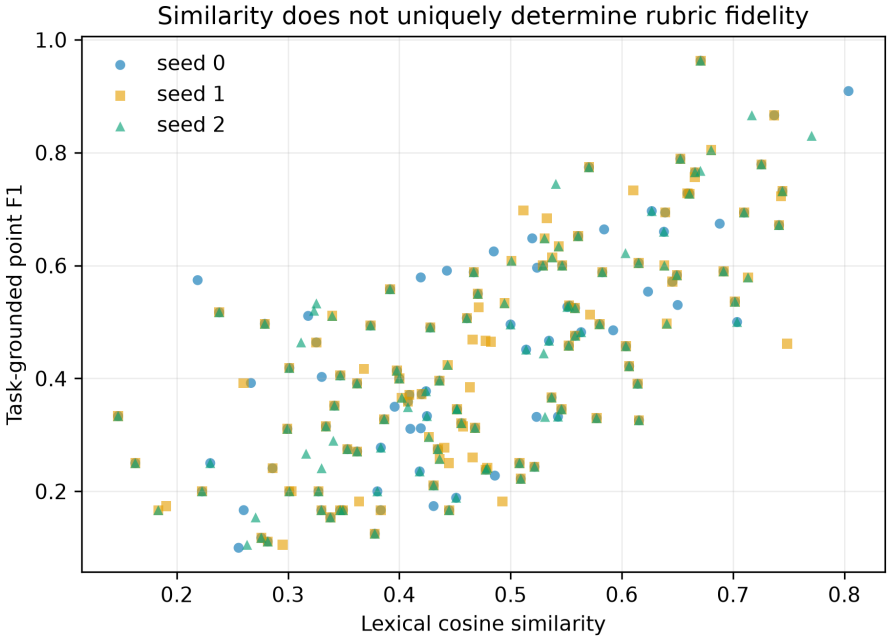


Fig. 1. Lexical cosine and constructed point-F1 for the 360 Qwen3-14B generations. The association is positive, but outputs with similar cosine scores can differ substantially in concept fidelity.

Reusing the language of a related training item increases surface resemblance but does not reliably transfer the correct support concepts. This divergence reinforces the need to inspect what each measure rewards.

4.2 Similarity and constructed concept fidelity

Figure 1 compares lexical cosine with point-F1 for all 360 Qwen3-14B outputs. Lexical cosine had Pearson correlation 0.704 with point-F1. BLEU and ROUGE-L correlations were 0.660 and 0.803, respectively. Reference resemblance is therefore informative on average in this sample.

The scatter also shows why a correlation is not a validity guarantee. At a given similarity level, point-F1 varies across items, and the three seeds largely overlap. More importantly, correlations over naturally generated outputs do not test whether a measure responds appropriately to a targeted polarity reversal or omission. That question is addressed by the controlled corruptions.

4.3 Sensitivity to consequential corruptions

Table 2 and Figure 2 report the corruption results. Omission, generic language, and unrelated support reduce all measures. Negation behaves differently. It

Table 2. Controlled corruption results. Values are means ± 95% confidence intervals over 120 items.

Variant	Point-F1	BLEU	ROUGE-L	Lexical cosine
Reference	0.814 ± 0.027	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Omission	0.452 ± 0.011	0.142 ± 0.059	0.427 ± 0.042	0.661 ± 0.027
Negation	0.724 ± 0.035	0.726 ± 0.032	0.849 ± 0.022	0.901 ± 0.016
Generic explanation	0.273 ± 0.012	0.145 ± 0.022	0.382 ± 0.023	0.545 ± 0.019
Unrelated support	0.301 ± 0.012	0.147 ± 0.026	0.434 ± 0.025	0.583 ± 0.020

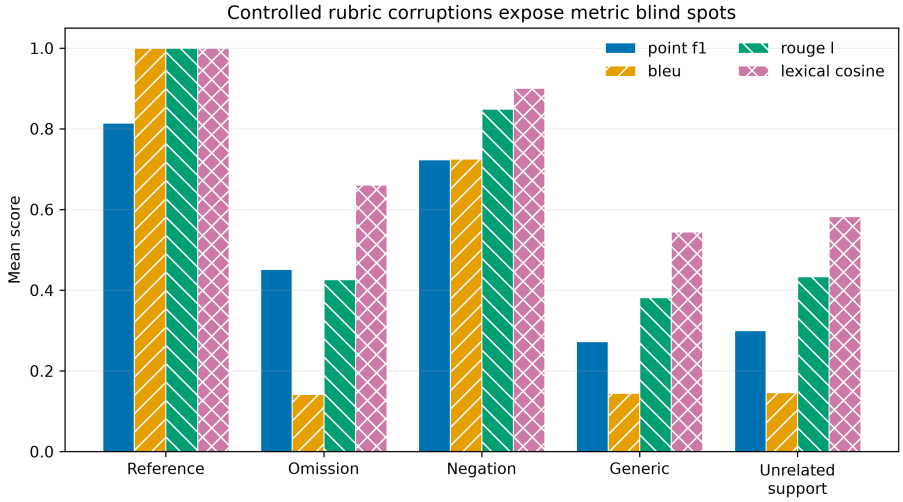


Fig. 2. Mean response of four automatic measures to controlled rubric corruptions. Negation preserves high scores because it retains most reference tokens while reversing the intended criterion.

retains most of the reference words and consequently receives BLEU 0.726, ROUGE-L 0.849, and lexical cosine 0.901. The implemented point-F1 also remains 0.724 because its token set does not represent polarity.

Negation exposes a shared limitation rather than a defect unique to one measure. N-gram overlap, sequence overlap, bag-of-token similarity, and the constructed concept metric all reward retained content words. None of them explicitly encodes whether a proposition is affirmed or denied. All four scores decline under negation, but they remain closer to the reference than the other tested corruptions. Without educator labels and an operational threshold, these values do not establish false acceptance. They instead show that a token-preserving polarity reversal receives comparatively high similarity under this protocol.

The other corruptions are easier for these measures because they replace or remove many content tokens. This is useful as a sanity check, but it also shows the narrowness of the stress suite. The unrelated-support corruption is a generic

off-target phrase, not a natural wrong mechanism. The omission corruption is degenerate for the 15 empty-support items because the constructed reference already contains no supporting explanation. These cases are retained so that the table matches the fixed 120-item protocol, but they should not be read as a validated distribution of rubric errors.

4.4 Complementary information from score association and corruption tests

The naturally generated outputs and deterministic corruptions answer different descriptive questions. Correlations over 360 outputs summarize whether similarity and constructed concept reproduction tend to move together in this sample. The paired transformations instead show how the same scores change when a reference is altered in a known way. Neither analysis supplies an educator-defined acceptance threshold, and the pooled correlations do not account for dependence among three outputs for each item. The evidence therefore supports a narrow conclusion: average association and response to a specific error should be reported separately.

This separation also prevents unlike score scales from being treated as interchangeable. Sentence BLEU, ROUGE-L, lexical cosine, and point-F1 have different ranges and reward structures in this experiment. Their absolute values cannot be interpreted as probabilities that a rubric is acceptable. Operational interpretation would require educator-labeled acceptable paraphrases and defective criteria, followed by threshold calibration on held-out items.

4.5 Error interpretation

The aggregate Qwen3-14B scores combine at least three kinds of behavior. Some outputs repeat the answer with little explanatory content. These cases tend to preserve answer-hit while losing support coverage. Other outputs supply a reasonable scientific explanation in vocabulary that differs from the support passage. The constructed point metric may penalize these paraphrases even when a teacher might accept them. A third group adds content that sounds scientific but is not anchored to the reference support. Without expert review, an automatic measure cannot reliably separate the latter two groups.

This ambiguity explains why the system ranking is secondary to the controlled transformations. The answer-only control is favored by brevity, while retrieval-edit benefits from copied wording. Qwen3-14B has more opportunity both to express valid paraphrases and to introduce unsupported content. Comparing these systems with one automatic target therefore mixes model quality with length and reference-style effects. The transformations avoid part of this problem because each corrupted output is paired with its own uncorrupted reference and differs in a prespecified way.

The corruption set is deliberately small but diagnostically distinct. Omission removes required evidence when support evidence exists; generic language replaces item-specific evidence with boilerplate; unrelated support introduces an

off-target placeholder; negation reverses the proposition while preserving its content words. The measures respond strongly to the first three because they change many tokens. Their weak response to negation is more concerning because educational scoring depends on the asserted relation, not simply the presence of the right terminology.

5 Discussion

5.1 What the results establish

The controlled corruptions provide more targeted evidence than the raw system ranking, but they are still proxy evidence. Under the constructed SciQ target, the open model usually states the answer but reproduces only part of the support language. The answer-only control scores higher on point-F1, and retrieval-edit scores higher on surface similarity than on constructed concept fidelity. These comparisons show that the measures reward different properties.

The negation test then supplies a direct counterexample to interpreting a high raw score as sufficient evidence of rubric validity. A polarity-reversed criterion should not receive credit in a grading protocol, yet it preserves comparatively high overlap and concept scores. This case does not estimate a false-acceptance rate, but it shows that positive correlation with a content metric does not by itself establish sharp separation of a consequential error.

This distinction matters for reporting. Aggregate correlations describe average ordering over the sampled outputs. Failure-oriented tests ask whether the evaluation behaves correctly on prespecified errors. Both are useful, but they answer different questions. For high-stakes assessment, the second question cannot be replaced by the first.

5.2 Implications for evaluation design

Generated-rubric studies should report at least three layers of evidence. The first is conventional reference comparison, including lexical and embedding-based measures. The second is criterion-level assessment: required concepts, forbidden claims, logical polarity, and item relevance. The third is human and downstream validation: educator judgments and agreement with teacher scoring on student responses. The present experiment addresses only lexical reference comparison and a proxy version of criterion checking.

Controls should also be chosen to expose what a model adds. An answer-only control reveals whether elaboration adds reference-grounded content. Retrieval-edit reveals whether surface resemblance can be achieved by borrowing language from a nearby item. For multimodal questions, text-only and diagram-only controls are necessary to show that the model uses visual evidence rather than textual shortcuts. Human-authored rubrics remain the appropriate criterion standard for usability and mark allocation.

Metric stress tests are inexpensive relative to model training, but they are not new as a general NLG-evaluation idea. Their value in this setting is that they

force the rubric-evaluation protocol to name the educational error it is expected to detect. Negation, omission, generic language, and unrelated support can be generated deterministically and inspected before any model comparison. Additional probes could alter quantities, entities, causal direction, mark allocation, or required conditions. A useful metric suite should respond monotonically to these changes when their severity is prespecified by educators.

The design also suggests a distinction between screening and validation. Similarity measures can screen a large output set for unusual cases, compare broad tendencies, or support sampling for human review. They should not be treated as validation of an individual criterion unless sensitivity to the relevant error classes has been established. For a classroom-facing system, a high average score cannot compensate for a small number of criteria that reverse the intended credit decision.

Threshold selection must therefore be based on intended use. A metric threshold chosen to maximize correlation with point-F1 may still admit negated criteria because the two measures share token-level information. A threshold intended for safety would instead be calibrated against educator-labeled failure cases, with separate reporting of false acceptance for polarity, omission, factual error, and irrelevance. The error distribution, not only the overall mean, determines whether review effort is manageable.

5.3 Reproducibility and research use

The public testbed is useful as a diagnostic component rather than a complete educational benchmark. The package records the split files and hashes, prompt configuration, seeds, endpoint metadata, raw generations, summary tables, and figure inputs. These materials allow another group to reproduce the automatic analysis and extend the corruption suite.

The same retained outputs can also be reanalyzed with embedding, learned, entailment, QA-based factuality, or LLM-judge metrics not evaluated here. In particular, future work can compute a fully specified BERTScore configuration and contrast its response with lexical measures on identical transformations. The important design requirement is to record the encoder, layer, tokenization, normalization, rescaling, reference aggregation, and software version so that the reported statistic has a stable interpretation.

Reproducibility is especially important when educational data cannot be redistributed. At minimum, studies can publish split rules, duplicate checks, metric code, prompt templates, decoding parameters, hashes of protected inputs and outputs, and a public proxy that exercises the same evaluation claim. These records do not replace access to authentic rubrics, but they make clear which findings derive from the model and which depend on an unreported implementation choice.

The present artifacts also permit several extensions without repeating model inference. Alternative tokenization, polarity-aware rules, entailment models, and embedding measures can be applied to the retained outputs. Researchers can add new corruptions to the same constructed references and compare sensitivity

curves. Such reuse turns the experiment into a measurement testbed even though it is not an adequate benchmark for educational effectiveness.

5.4 Requirements for educational validation

The next stage should replace constructed support references with teacher-written marking criteria and explicitly measure agreement among educators. Each output should be rated for scientific correctness, criterion completeness, relevance, clarity, polarity, mark allocation, and estimated edit effort. Multiple raters are necessary because legitimate rubrics may differ in granularity and wording. Adjudicated labels can then be used to test whether automatic measures rank genuinely acceptable paraphrases above fluent but defective criteria.

Downstream scoring provides a second validation layer. Generated and human rubrics should be applied to student responses, with teacher scores serving as the criterion. Agreement, calibration, and disagreement cases would show whether an apparently small rubric error changes actual credit decisions. Where demographic or contextual subgroup analysis is lawful and sufficiently powered, it should be reported because aggregate agreement can conceal uneven error patterns.

Multimodal claims require a separate design. A benchmark should label which questions require a diagram, then compare text-only, diagram-only, and multimodal conditions on the same items. Performance on questions solvable from the text does not establish visual grounding. Diagram perturbations and counterfactual image-text pairs can further test whether a model uses the depicted evidence rather than question-stem regularities.

A broader model comparison would include several open instruction models, a task-specific or fine-tuned generator, extraction and retrieval controls, and human-authored references. Prompt and decoding variation should be prespecified, and uncertainty should be reported across both items and settings. These additions would allow conclusions about model classes; the current single-model experiment supports conclusions only about the evaluated protocol.

6 Limitations

The references are deterministic transformations of SciQ support passages, not educator-authored rubrics. Consequently, point-F1 measures constructed concept reproduction rather than teaching quality, rubric completeness, mark allocation, or consistency in actual grading. The prompt gives the correct answer to the generator, while the support passage used for evaluation is withheld; answer-hit is therefore a prompt-adherence check and support coverage partly reflects an information-access mismatch. The evaluation contains the first 120 text-only test items from one dataset and one open-weight model; item selection was not randomized or stratified, 15 selected items have empty support, and duplicate or

pretraining contamination was not assessed. It does not evaluate diagram comprehension, hosted or fine-tuned models, contextual embedding metrics, learned metrics, entailment or QA factuality metrics, LLM judges, or transfer to another science-question source. The three decoding seeds quantify a narrow source of generation variability but do not cover prompt variation, model variants, model families, or inference-engine revisions. The pooled metric correlations treat repeated outputs for an item as separate observations and are descriptive rather than hierarchical uncertainty estimates.

Lexical cosine is a fully specified reference-similarity measure and is not equivalent to BERTScore. Point-F1 is polarity-blind and is not normalized to 1.000 for the template reference. The deterministic corruptions are illustrative probes rather than a validated sample of realistic model errors; they lack educator severity labels, matched acceptable paraphrases, and an operational acceptance threshold. Omission is identical to the constructed reference on empty-support items, and unrelated support is a placeholder phrase rather than a plausible misconception. No science educators rated the outputs, and the study does not measure downstream scoring of student responses, educator edit effort, calibration, subgroup behavior, or classroom effects. These limitations restrict the findings to automatic evaluation behavior under the stated proxy protocol.

7 Conclusion

Reference similarity is useful but insufficient for evaluating generated science rubrics. In the SciQ proxy, Qwen3-14B produced stable answer-containing outputs across seeds, yet a concise answer-only control achieved higher constructed point-F1. Polarity-reversed references retained comparatively high lexical cosine, ROUGE-L, and point-F1 scores, although all declined from the unmodified reference. This diagnostic case study does not establish an operational false-acceptance rate. It shows that reference-similarity values should not stand alone as proof of rubric validity. Generated-rubric studies should combine paraphrase-tolerant similarity, explicit criterion checks, controlled failure tests, educator ratings, and downstream grading evaluation.

Data availability

SciQ is publicly available from the Allen Institute for AI at `https://huggingface.co/datasets/allenai/sciq` under the CC BY-NC 3.0 license and is described by **(author?)** [40]. Exact split URLs, byte sizes, row counts, and SHA256 hashes are provided in `DATA.md`. Derived result tables and raw model generations are included with the analysis package. No student records or private examination data were used.

Code availability

The code used for the reported analysis is included with the review package. A public archival release will be provided at publication.

CRediT author contribution

Lauren Mercer, Maya Kalyani Reed, Hoang Justin Tran, Alexander J. Garnier: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration.

Funding

This research received no external funding.

Declaration of competing interest

The author declares no competing interests.

Declaration of generative AI use

During the preparation of this work the author used AI-assisted tools for literature search, code development and drafting. The author reviewed and edited the content and takes full responsibility for it.

Bibliography

[1] Abdul Salam, M., El-Fatah, M.A., Hassan, N.F.: Automatic grading for arabic short answer questions using optimized deep learning model. PLOS ONE **17**(8), e0272269 (Aug 2022). https://doi.org/10.1371/journal.pone.0272269, http://dx.doi.org/10.1371/journal.pone.0272269

[2] Ahmed, A., Joorabchi, A., Hayes, M.: On deep learning approaches to automated assessment: Strategies for short answer grading. In: Proceedings of the 14th International Conference on Computer Supported Education. pp. 85–94. SCITEPRESS - Science and Technology Publications (2022). https://doi.org/10.5220/0011082100003182, http://dx.doi.org/10.5220/0011082100003182

[3] Ariely, M., Nazaretsky, T., Alexandron, G.: Machine learning and hebrew nlp for automated assessment of open-ended questions in biology. International Journal of Artificial Intelligence in Education **33**(1), 1–34 (Mar 2023). https://doi.org/10.1007/s40593-021-00283-x, http://dx.doi.org/10.1007/s40593-021-00283-x

[4] Arthur Peterson, K.: Leveraging deep learning for automated generative grading of science subject-based structured questions. Preprints (Oct 2026). https://doi.org/10.20944/preprints202511.0699.v2, http://dx.doi.org/10.20944/preprints202511.0699.v2

[5] Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education **25**(1), 60–117 (Mar 2015). https://doi.org/10.1007/s40593-014-0026-8, http://dx.doi.org/10.1007/s40593-014-0026-8

[6] Caglayan, O., Sanabria, R., Palaskar, S., Barraul, L., Metze, F.: Multimodal grounding for sequence-to-sequence speech recognition. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8648–8652. IEEE (May 2019). https://doi.org/10.1109/icassp.2019.8682750, http://dx.doi.org/10.1109/icassp.2019.8682750

[7] Chung, M.Y., Fan, J., Ye, H., Wang, Q., Shen, W.C., Yu, C.M., Chen, P.Y., Kuo, S.Y.: Model reprogramming demystified: A neural tangent kernel perspective. arXiv preprint arXiv:2506.00620 (2025), https://arxiv.org/abs/2506.00620

[8] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423, http://dx.doi.org/10.18653/v1/n19-1423

[9] EL Mernissi, A., Lakrami, F., Toumi, H.: A comparative analytical study of llm-based automated grading in programming education. Social Sciences &

Humanities Open **14**, 103144 (Dec 2026). https://doi.org/10.1016/j.ssaho.2026.103144, http://dx.doi.org/10.1016/j.ssaho.2026.103144

[10] Eshetu, A., Teshome, G., Alemahu, R.: Bi-directional answer-to-answer co-attention for short answer grading using deep learning. In: Proceedings of the The Fourth Widening Natural Language Processing Workshop. pp. 26–30. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.winlp-1.8, http://dx.doi.org/10.18653/v1/2020.winlp-1.8

[11] Fan, J., Liang, Y., Liu, L., Huang, S., Zhang, L.: Rca-noc: Relative contrastive alignment for novel object captioning. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15464–15474. IEEE (Oct 2023). https://doi.org/10.1109/iccv51070.2023.01422, http://dx.doi.org/10.1109/iccv51070.2023.01422

[12] Fan, J., Rosu, P., Wang, A.T., Li, Z.M., Carin, L., Cheng, X.: In-context semi-supervised learning. arXiv preprint arXiv:2512.15934 (2025), https://arxiv.org/abs/2512.15934

[13] Ghavidel, H., Zouaq, A., Desmarais, M.: Using bert and xlnet for the automatic short answer grading task. In: Proceedings of the 12th International Conference on Computer Supported Education. pp. 58–67. SCITEPRESS - Science and Technology Publications (2020). https://doi.org/10.5220/0009422400580067, http://dx.doi.org/10.5220/0009422400580067

[14] Haller, S., Aldea, A., Seifert, C., Strisciuglio, N.: Survey on automated short answer grading with deep learning: from word embeddings to transformers. arXiv preprint arXiv:2204.03503 (2022), https://arxiv.org/abs/2204.03503

[15] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE (June 2016). https://doi.org/10.1109/cvpr.2016.90, http://dx.doi.org/10.1109/CVPR.2016.90

[16] Kazi, N., Kahanda, I.: Enhancing transfer learning of llms through fine- tuning on task - related corpora for automated short-answer grading. In: 2023 International Conference on Machine Learning and Applications (ICMLA). pp. 1687–1691. IEEE (Dec 2023). https://doi.org/10.1109/icmla58977.2023.00255, http://dx.doi.org/10.1109/icmla58977.2023.00255

[17] Kour, G., Ackerman, S., Farchi, E.D., Raz, O., Carmeli, B., Tavor, A.A.: Measuring the measuring tools: An automatic evaluation of semantic metrics for text corpora. In: Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM). pp. 405–416. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.gem-1.35, http://dx.doi.org/10.18653/v1/2022.gem-1.35

[18] Lam, H.W., Dillon, T., Chang, E.: Determining writing genre: Towards a rubric-based approach to automated essay grading. In: 2011 IEEE International Conference on Advanced Information Networking and Applications. pp. 270–274. IEEE (Mar 2011). https://doi.org/10.1109/aina.2011.32, http://dx.doi.org/10.1109/aina.2011.32

[19] Leung, W.C., Wein, S., Schneider, N.: Semantic similarity as a window into vector- and graph-based metrics. In: Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM). pp. 106–115. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.gem-1.8, http://dx.doi.org/10.18653/v1/2022.gem-1.8

[20] Liang, Y., Yang, Z., Tang, Y., Fan, J., Li, Z., Wang, J., Torr, P.H., Huang, S.L.: Luna: Language as continuing anchors for referring expression comprehension. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5174–5184. MM '23, ACM (Oct 2023). https://doi.org/10.1145/3581783.3612584, http://dx.doi.org/10.1145/3581783.3612584

[21] Mahmoud, S., Torki, M.: Alexu-aux-bert at semeval-2020 task 3: Improving bert contextual similarity using multiple auxiliary contexts. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 270–274. International Committee for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.semeval-1.33, http://dx.doi.org/10.18653/v1/2020.semeval-1.33

[22] Manoharan, S.: A framework for automated assignment generation and marking for plagiarism mitigation. In: 2017 27th EAEEIE Annual Conference (EAEEIE). pp. 1–6. IEEE (June 2017). https://doi.org/10.1109/eaeeie.2017.8768680, http://dx.doi.org/10.1109/eaeeie.2017.8768680

[23] Nakov, P., Màrquez, L., Magdy, W., Moschitti, A., Glass, J., Randeree, B.: Semeval-2015 task 3: Answer selection in community question answering. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 269–281. Association for Computational Linguistics (2015). https://doi.org/10.18653/v1/s15-2047, http://dx.doi.org/10.18653/v1/s15-2047

[24] Ni'mah, I., Fang, M., Menkovski, V., Pechenizkiy, M.: Nlg evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1240–1266. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.acl-long.69, https://aclanthology.org/2023.acl-long.69/

[25] Papachristou, I., Dimitroulakos, G., Vassilakis, C.: Automated test generation and marking using llms. Electronics **14**(14), 2835 (July 2025). https://doi.org/10.3390/electronics14142835, http://dx.doi.org/10.3390/electronics14142835

[26] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL 02. p. 311. ACL 02, Association for Computational Linguistics (2001). https://doi.org/10.3115/1073083.1073135, http://dx.doi.org/10.3115/1073083.1073135

[27] Peris, A., Casacuberta, F.: A neural, interactive-predictive system for multimodal sequence to sequence tasks. In: Proceedings of the 57th Annual

Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 81–86. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/p19-3014, http://dx.doi.org/10.18653/v1/p19-3014

[28] Pham, H., Manzini, T., Liang, P.P., Poczós, B.: Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. In: Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML). pp. 53–63. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/w18-3308, http://dx.doi.org/10.18653/v1/w18-3308

[29] Putnikovic, M., Jovanovic, J.: Embeddings for automatic short answer grading: A scoping review. IEEE Transactions on Learning Technologies **16**(2), 219–231 (Apr 2023). https://doi.org/10.1109/tlt.2023.3253071, http://dx.doi.org/10.1109/TLT.2023.3253071

[30] Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S.: Beyond accuracy: Behavioral testing of nlp models with checklist. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4902–4912. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.442, https://aclanthology.org/2020.acl-main.442/

[31] Rudzewitz, B., Ziai, R.: Comic: Adapting a short answer assessment system for answer selection. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 247–251. Association for Computational Linguistics (2015). https://doi.org/10.18653/v1/s15-2044, http://dx.doi.org/10.18653/v1/s15-2044

[32] Rus, V.: Explanation-based automated assessment of open ended learner responses. In: 14th International Conference eLearning and Software for Education. eLSE, vol. 2, pp. 120–127. Carol I National Defence University Publishing House (Apr 2018). https://doi.org/10.12753/2066-026x-18-087, http://dx.doi.org/10.12753/2066-026x-18-087

[33] Saadany, H., Orasan, C.: Bleu, meteor, bertscore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. In: Proceedings of the Translation and Interpreting Technology Online Conference TRITON 2021. pp. 48–56. TRITON 2021, INCOMA Ltd. Shoumen, BULGARIA (2021). https://doi.org/10.26615/978-954-452-071-7_006, http://dx.doi.org/10.26615/978-954-452-071-7_006

[34] Sai, A.B., Dixit, T., Sheth, D.Y., Mohan, S., Khapra, M.M.: Perturbation checklists for evaluating nlg evaluation metrics. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 7219–7234. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.575, https://aclanthology.org/2021.emnlp-main.575/

[35] Satcharattanachot, P., Usanavasin, S.: Automated grading approach for open-ended stem answers using llm. In: 2025 20th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP). pp. 1–6. IEEE (Nov 2025). https://doi.org/10.1109/isai-nlp

66160.2025.11320498, http://dx.doi.org/10.1109/isai-nlp66160.2025.11320498

[36] Shehab, A., Elhoseny, M., Hassanien, A.E.: A hybrid scheme for automated essay grading based on lvq and nlp techniques. In: 2016 12th International Computer Engineering Conference (ICENCO). pp. 65–70. IEEE (Dec 2016). https://doi.org/10.1109/icenco.2016.7856447, http://dx.doi.org/10.1109/icenco.2016.7856447

[37] Skalnes, E., Price, L.C., Anantha, R., Oberst, M.: Evaluating rubric generation with interventional transfer. arXiv preprint arXiv:2610.10809 (2026), https://arxiv.org/abs/2610.10809

[38] Sultan, M.A., Salazar, C., Sumner, T.: Fast and easy short answer grading with high accuracy. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1070–1075. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/n16-1123, http://dx.doi.org/10.18653/v1/N16-1123

[39] Tobler, S.: Smart grading: A generative ai-based tool for knowledge-grounded answer evaluation in educational assessments. MethodsX **12**, 102531 (June 2024). https://doi.org/10.1016/j.mex.2023.102531, http://dx.doi.org/10.1016/j.mex.2023.102531

[40] Welbl, J., Liu, N.F., Gardner, M.: Crowdsourcing multiple choice science questions. arXiv preprint arXiv:1707.06209 (2017), https://arxiv.org/abs/1707.06209

[41] Wieting, J., Berg-Kirkpatrick, T., Gimpel, K., Neubig, G.: Beyond bleu: Training neural machine translation with semantic similarity. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4344–4355. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/p19-1427, http://dx.doi.org/10.18653/v1/p19-1427

[42] Wijanto, M.C., Karnalim, O., Tan, R.: Work in progress: High school students' perspective on assessment question types during online learning - preliminary study for automated assessments of open-ended questions. In: 2022 IEEE World Engineering Education Conference (EDUNINE). pp. 1–4. IEEE (Mar 2022). https://doi.org/10.1109/edunine53672.2022.9782327, http://dx.doi.org/10.1109/edunine53672.2022.9782327

[43] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025), https://arxiv.org/abs/2505.09388

[44] Yu, T., Ekbote, C., Morozov, N., Herrera, A., Fan, J., Dominguez Mantes, A., Novello, S., Lashuel, H., Frossard, P., d'Ascoli, S., La Manno, G., Brbic,

M.: Tissue reconstruction with generative ai. bioRxiv (Feb 2025). https://doi.org/10.1101/2025.02.13.638045, http://dx.doi.org/10.1101/2025.02.13.638045

[45] Zeidler, L., Opitz, J., Frank, A.: A dynamic, interpreted checklist for meaning-oriented nlg metric evaluation – through the lens of semantic similarity rating. In: Proceedings of the 11th Joint Conference on Lexical and Computational Semantics. pp. 157–172. Association for Computational Linguistics, Seattle, Washington (2022). https://doi.org/10.18653/v1/2022.starsem-1.14, https://aclanthology.org/2022.starsem-1.14/

[46] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019), https://arxiv.org/abs/1904.09675

[47] Zhou, J., Zhang, Q., Wang, Y., Lyu, F., Ming, Y., Xu, C., Sun, Q., Zheng, K., Kang, P., Liu, X., Ma, C.: Rubricbench: Aligning model-generated rubrics with human standards. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 31179–31200. Association for Computational Linguistics, San Diego, California, United States (2026). https://doi.org/10.18653/v1/2026.acl-long.1439, https://aclanthology.org/2026.acl-long.1439/

[48] Zhu, H., Li, T., He, P., Zhou, J.: Enhancing automated grading in science education through llm-driven causal reasoning and multimodal analysis. In: Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence. pp. 10352–10360. IJCAI-2024, International Joint Conferences on Artificial Intelligence Organization (Aug 2024). https://doi.org/10.24963/ijcai.2024/1150, http://dx.doi.org/10.24963/ijcai.2024/1150