

Evaluating Multi-Phase and Granular Strategies for Evidence-Anchored Automated Scoring

Alessia Marigo Laura Wright Linda Malkin Xin Xie
amarigo@wisc.edu laura.wright@wisc.edu lmmalkin@wisc.edu xin.xie@wisc.edu

School of Education, University of Wisconsin–Madison
1025 W Johnson St, Madison, WI 53706-1706

Abstract

This study evaluates few-shot large language models (LLMs) on middle-school geoscience responses ($N=86$ responses $\times$ 11 indicators), separating *presence* agreement from *extract* agreement. Joint prompting and annotation-like rubric guidance plus examples yield the clearest gains; lengthy rubric rewriting and clause-level parsing do not reliably improve extract alignment.

1 Introduction

Large language models (LLMs) can score assessments with high internal reliability (Yavuz et al., 2025), yet accuracy sufficient to generate personalized formative feedback remains challenging (Wisniewski et al., 2020). This is especially important in science education, where multidimensional standards target growth across ideas, practices, and crosscutting concepts (NGSS Lead States, 2013). Effective feedback must speak to each dimension separately.

Mixed LLM pipelines and chain-of-thought scoring (Lee et al., 2024), evidence-centered design for AI scoring (Ercikan and McCaffrey, 2022), multi-agent component extraction (Wang et al., 2026), consistency analyses of LLM scorers (Xue et al., 2026), and rubric-verbosity / inference-level effects (Kucia et al., 2026) motivate our experiments. Annotation-aware essay scoring (Ormerod, 2025) further motivates testing finer response units (clauses), which we evaluate and ultimately do not retain as a primary method.

In prior work (Marigo et al., 2026), fine-tuned Llama 3 showed high reliability ($\alpha=0.92$) with human raters, while Gemini was competitive on presence κ but weak on Fuzzy κ for spans. Both families could often detect indicator presence yet failed to recover the human consensus evidence span (verbatim snippet). For feedback, the model must be accurate in *evidence anchoring*, not only in presence

classification. This paper maps which changes—architecture, rubric engineering, or granularity—actually move that needle.

2 Aims

- **System architecture:** Does a dual-phase detect-then-extract pipeline outperform single-phase end-to-end scoring for presence and for verbatim evidence?
- **Rubric engineering:** Which rubric revisions (error polarity, oracle-informed language, additional shots) improve agreement across indicators?
- **Granularity:** Does clause-level parsing improve evidence alignment with the human consensus (oracle)?

3 Sample

Data come from Grade 7–8 science students in the ONPAR project (2019–2020) (Wright et al., 2022). Four teachers piloted an earth-science writing task (two suburban Midwest; two rural East Coast). Two expert raters developed an 11-indicator three-dimensional science rubric and double-coded 100 responses with a *span* or *none* per indicator, then formed a consensus set.

3.1 Writing prompt

The writing prompt provided students with a map of North and South America, tectonic plates in the region, and a series of red dots indicating earthquakes. In addition, a data table provided information about plate movement, including the speed and direction of plate movement. Students were asked to study the information to identify a pattern and make a conclusion about plate movement and earthquakes (see Figure 1).

Writing Prompt
The image below shows major earthquakes in a small area of Earth. Study the image and read the data on plate movement. What does the pattern of data allow you to conclude about plate movement and earthquakes?

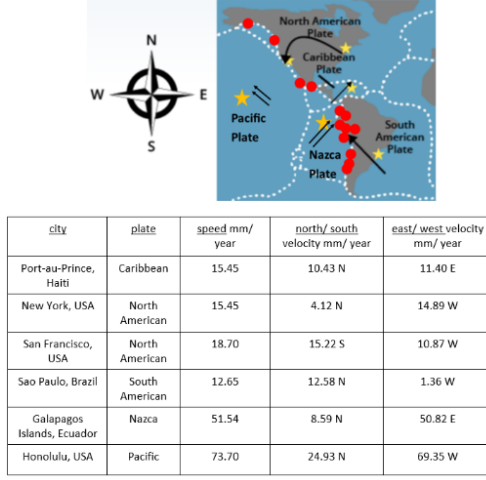


Figure 1: ONPAR earth-science constructed-response task: map and table stimuli with the student writing prompt.

3.2 Rubric indicators

The rubric accounts for the three dimensions of the NGSS (NGSS Lead States, 2013): Disciplinary Core Ideas (DCI), Science and Engineering Practices (SEP), and Crosscutting Concepts (CCC). Table 1 lists the 11 indicators used in scoring (canonical names under polarity-inverted rubrics B/C; rubric A uses error-present wording for the four error/absence indicators).

3.3 Coding and analysis sample

Of the 100 consensus-coded responses, some were reserved as few-shot examples and some were dropped for near-total *none* coding. The analysis set—hereafter the *oracle*—is **86 responses** $\times$ **11 indicators** (946 cells).

4 Methodology

4.1 Measures

Presence is binary: span = 1, none = 0. Let $[a, b]$ be the *combined presence codes* for model and oracle.

When both are present ($[1, 1]$ combined presence code), we score span quality. Two spans are a **76% match** if *token-sort* similarity is ≥ 76 (not raw character overlap). We compute this using RapidFuzz (Bachmann, 2025), which splits each span into individual words (tokens) and sorts them alphabetically before calculating their similarity score.

This ensures that minor variations in word order or punctuation do not artificially penalize the match. For end-to-end Fuzzy-76 κ we use *combined span codes* (a, b) where:

- $(0, 0)$: both have *none*;
- $(0, 1)$: oracle has a *span* and the model is either *none* or not a 76% match;
- $(1, 0)$: only the model returns a *span*;
- $(1, 1)$: model and oracle both return a *span* with a 76% match.

Prior work yielded usable presence κ but weak Fuzzy-76 κ . We therefore re-ran with the strongest prior configurations—Gemini 3.1-pro (few-shot) and Llama-3.1-8B-Instruct LoRA ($r=64$, $\alpha=16$, 5-fold)—and confirmed the gap (Table 2).

4.2 Strategies

4.2.1 Two-phase coding

We separate *detect* (0/1 per indicator) from *extract* (verbatim span only if detect=1). For extract quality we report the agreement rate among eligible cells, that is both oracle and detect return =1:

$$76\% \text{ Rate} = \frac{\#(1, 1)}{\#(1, 1) + \#(0, 1)}, \quad (1)$$

where here $\#(1, 1)$ means 76% match and $\#(0, 1)$ means a mismatch among eligible cells (combined presence code = $[1, 1]$).

4.2.2 Additional measures

Many cells are *none*, partly because several indicators are error/absence constructs: even a strong response yields several zeros, producing imbalance. Cohen’s κ is known to be deflated under skewed marginals; indicator-level κ ranged roughly from 0.10 to 0.95. Gwet’s AC1 is more stable under imbalance (illustrative range ≈ 0.58 –0.95 where paired κ was lower).

4.2.3 Rubric revisions

- **A:** original rubric (error indicators scored as error-present).
- **B:** A with inverted error polarity (absence-of-error framed as the positive; Table 1).
- **C:** B plus oracle-informed indicator language (what to quote / not quote / when none) and additional paraphrased shots targeting hard extract indicators.

ID	Dim.	Focus (consensus evidence)
R1_INTERACTION	DCI	Two plates interacting; boundaries, borders, or edges
R1_MOVE	DCI	Plate motion (converge/push/collide/slip/transform; relative direction or speed)
R1_HAZARD	DCI	Explicit causal link: earthquakes caused by plate movement
R1_NO_MINOR_ERROR	DCI	Absence of minor motion-wording / out-of-scope error (else quote the error)
R1_NO_MAJOR_ERROR	DCI	Absence of major wrong-motion / wrong-cause error (else quote the error)
R2_DATA_MAP	SEP	Relevant map features (red dots, boundaries, arrows, named plates)
R2_DATA_TABLE	SEP	Relevant table speeds/directions (esp. Nazca / South America)
R2_DATA_INTERPRET	SEP	Correct interpretation or comparison of identified data
R2_NO_INCORRECT_DATA	SEP	Absence of incorrect map/table reading (else quote the error)
R3_PATTERN_IMPLICIT	CCC	Meaningful pattern without an explicit quantifier
R3_PATTERN_EXPLICIT	CCC	Meaningful pattern with explicit quantifier (most/all/preponderance, etc.)

Table 1: Eleven NGSS-aligned scoring indicators. Dimension follows the ONPAR R1/R2/R3 grouping (DCI / SEP / CCC). Brief focus summarizes what consensus raters treat as positive evidence.

Model	Cohen κ	Fuzzy 76% κ
Gemini	0.52	0.25
Llama	0.32	0.07

Table 2: Baseline presence vs. Fuzzy-76 agreement (means over indicators).

4.2.4 Loop by indicator

We also run joint (all indicators in one call) vs. indicator-looped calls for both one-phase and two-phase designs.

4.2.5 Other strategies (tested, not retained)

Since Llama showed frequent hallucinations on this task, we reported arms use Gemini 3.1-pro only. A high-recall detect patch was attempted and while it reduced misses, the number of false alarms exploded. Also a clause-level extract (select or aggregate clauses) was tried but it did not beat thin free-span extract; see Results (Aim 3).

4.2.6 Cont&76% Rate diagnostic

Many mismatches are incomplete but topical: the model span is often a substring (or token subset) of the oracle claim ($\text{LLM} \subset \text{oracle} \approx 50\text{--}60\%$ of Fuzzy-76 mismatches on joint arms), while oracle $\subset$ LLM is rarer ($\approx 10\text{--}15\%$). We therefore also report

$$\text{Cont\&76\% Rate} = \frac{\# \left\{ \begin{array}{l} \text{76\% match} \\ \text{or bidirectional} \\ \text{containment} \end{array} \right\}}{\#\{\text{eligible cells}\}}, \quad (2)$$

where eligible cells are those with combined presence code = [1, 1], as a *diagnostic* of grain (not as the operational scoring rule). Arm-level Cont&76% Rate and 76% Rate are **pooled** over eligible cells.

4.3 Experimental grid

Gemini 3.1-pro; rubrics A/B/C; architectures:

- 0 one-phase, joint;
- 1 one-phase, indicator-looped;
- 2 two-phase detect→extract, joint;
- 3 two-phase, indicator-looped.

Arm names combine rubric and architecture (e.g., A0, C2).

4.4 Statistical power

With $N=86$ responses as the clustering unit (946 cells are not independent), minimum detectable differences for paired arm contrasts are on the order of $\approx 12\text{--}15$ percentage points for 76% Rate and $\approx 5\text{--}12$ for presence agreement (80% power, $\alpha=0.05$, moderate within-response correlation). The size of our dataset is large enough for major metric shifts such as Cont&76% Rate vs. 76% Rate ($\approx +30$ pp) and some architecture contrasts such as joint rubric vs indicator-looped. It is underpowered for extract improvements based on fine rubric rewording with observed mean deltas of 2-8 pp on 76% Rate.

5 Results

Mean differences across arms in Table 3 look modest, but indicator-level spread is large. Table 4

Arm	Description	Cohen κ	Gwet AC1	76% Rate	Cont&76% Rate
A0	Joint Rubric A 1-Phase	0.528	0.700	0.469	0.823
A1	IndLoop Rubric A 1-Phase	0.466	0.495	0.462	0.817
A2	Joint Rubric A 2-Phases	0.548	0.705	0.441	0.795
A3	IndLoop Rubric A 2-Phases	0.455	0.491	0.492	0.805
B0	Joint Rubric B 1-Phase	0.540	0.698	0.443	0.814
B1	IndLoop Rubric B 1-Phase	0.471	0.524	0.451	0.804
B2	Joint Rubric B 2-Phases	0.547	0.701	0.455	0.809
B3	IndLoop Rubric B 2-Phases	0.472	0.526	0.475	0.792
C0	Joint Rubric C 1-Phase	0.509	0.710	0.485	0.818
C1	IndLoop Rubric C 1-Phase	0.530	0.600	0.502	0.779
C2	Joint Rubric C 2-Phases	0.522	0.720	0.479	0.799
C3	IndLoop Rubric C 2-Phases	0.517	0.578	0.515	0.786

Table 3: Presence and extract metrics by arm. Cohen κ and Gwet AC1 are means over 11 indicators. 76% Rate and Cont&76% Rate are *pooled* over eligible cells (oracle present $\cap$ detect=1).

reports the same pooled Cont&76% Rate as Table 3, with min/max across the 11 indicators (AC1 and Cont&76% Rate).

5.1 Hard indicators and annotation priorities

Presence can look strong while extract fails. Averaging 76% Rate across joint arms A0/B0/C0, the weakest extract indicators are R1_INTERACTION (≈ 0.22), R1_NO_MAJOR_ERROR (≈ 0.35), R1_MOVE (≈ 0.38), R2_DATA_INTERPRET (≈ 0.39), and R3_PATTERN_EXPLICIT (≈ 0.42). By contrast, R1_HAZARD and R2_DATA_TABLE often exceed ≈ 0.75 – 0.80 on 76% Rate. R2_DATA_MAP is mixed: presence is moderate, but mismatches are frequently bare plate names instead of the full location claim (high Cont&76%, low 76% Rate on several arms).

For future rubric *annotations*, priority should go to:

- **R1_INTERACTION / R1_MOVE:** require the full interaction or relative-motion claim; forbid quoting only “edges,” “boundaries,” or bare speeds.
- **R2_DATA_MAP vs. R2_DATA_INTERPRET:** location/map features vs. comparison/meaning; forbid reusing one span for both.
- **R3_PATTERN_EXPLICIT:** quote the pattern-conclusion sentence (with quantifier), not only supporting detail.
- **R1_NO_MAJOR_ERROR:** when an error is present, quote only the erroneous

causal/motion claim (grain still unstable).

R1_HAZARD and R2_DATA_TABLE need less annotation effort under current defs.

5.2 Answers to the aims

Aim 1 (architecture). Dual-phase detect $\rightarrow$ extract does *not* clearly outperform one-phase on 76% Rate once detection errors remain (e.g., A2 0.441 vs. A0 0.469). Joint prompting (arms *0/*2) is more stable for **presence** (AC1 ≈ 0.70) than indicator-looped prompting (arms *1/*3), where mean AC1 often falls to ≈ 0.50 and some indicators approach zero or negative AC1. Two-phase helps mainly as an analytic separation of detect vs. extract error—not as an automatic accuracy win on mean 76% Rate.

Aim 2 (rubric engineering). Polarity inversion (A $\rightarrow$ B) clarifies error indicators under AC1 more than it lifts mean 76% Rate. Oracle-informed language plus paraphrased hard-indicator shots (B $\rightarrow$ C) yields modest extract gains on some looped arms (C1/C3 76% Rate) and competitive joint presence AC1 (C2 0.720), without a large uniform mean presence lift. Annotation-like guidance and examples matter more than lengthy rewrite alone; the hard indicators in Table 5 remain the main extract bottleneck.

Aim 3 (granularity). Clause-level extraction did not improve evidence alignment. Consensus spans are typically full claims (~ 90 – 100 characters). Clause-*selection* mismatches remained

Arm	AC1	Min	Max	Cont&76% Rate	Min	Max
A0	0.700	0.489	0.886	0.823	0.462	1.000
A1	0.495	-0.061	0.844	0.817	0.562	1.000
A2	0.705	0.465	0.910	0.795	0.615	0.914
A3	0.491	-0.015	0.826	0.805	0.500	0.970
B0	0.698	0.526	0.909	0.814	0.471	0.971
B1	0.524	-0.061	0.844	0.804	0.562	0.941
B2	0.701	0.349	0.926	0.809	0.500	0.938
B3	0.526	-0.086	0.837	0.792	0.500	0.935
C0	0.710	0.481	0.886	0.818	0.583	1.000
C1	0.600	0.090	0.930	0.779	0.500	0.917
C2	0.720	0.568	0.885	0.799	0.600	0.930
C3	0.578	0.008	0.930	0.786	0.556	0.953

Table 4: Spread across indicators. Gwet AC1 columns are the mean / min / max over indicators. Cont&76% Rate is the *pooled* arm rate (identical to Table 3), with min/max of per-indicator Cont&76% Rate.

Indicator	76% Rate	Cont&76%	AC1
R1_INTERACTION	0.222	0.814	0.877
R1_NO_MAJOR_ERROR	0.350	0.850	0.732
R1_MOVE	0.382	0.764	0.622
R2_DATA_INTERPRET	0.393	0.505	0.735
R3_PATTERN_EXPLICIT	0.421	0.724	0.533
R2_DATA_MAP	0.459	0.909	0.622
R3_PATTERN_IMPLICIT	0.522	0.767	0.544
R2_NO_INCORRECT_DATA	0.548	0.865	0.583
R1_NO_MINOR_ERROR	0.710	0.952	0.706
R1_HAZARD	0.757	0.941	0.886
R2_DATA_TABLE	0.802	0.844	0.887

Table 5: Mean 76% Rate / Cont&76% Rate / presence AC1 by indicator on joint arms A0, B0, and C0 (unweighted mean of the three arms). Lower 76% Rate = higher priority for claim-level rubric annotations and contrastive shots.

mostly $\text{LLM} \subset \text{oracle}$ (under-quoting), matching free-span grain errors. Clause-*aggregation* flipped errors toward $\text{oracle} \subset \text{LLM}$ (over-merging) and lowered 76% Rate. Clause boundaries are not the same unit humans quote; Cont&76% Rate rises because many Fuzzy-76 “misses” are contained fragments of the oracle. The clause-level extraction helped us diagnose the grain gap and the Cont&76% Rate was key to not underestimate working extraction methods.

6 Discussion

Our results offer key insights for rubric design and model instruction. Recent work shows model-written rubrics underperform human standards in rubric-guided evaluation (Zhou et al., 2026); our results similarly favor expert-calibrated indicator language over extensive automated rewrite. Sim-

ply expanding a rubric or flipping its polarity does not guarantee automatic gains in extract accuracy. Instead, targeted, expert-informed instructions and precise examples (such as those in Rubric C) are proven to be more effective in addressing specific extraction bottlenecks.

For evidence-anchored formative scoring, we recommend reporting presence and extraction agreement separately, favoring AC1 (or similar) alongside κ under imbalance, and treating Cont&76% Rate as a sensitivity check on grain. Operational systems should prioritize joint presence prompts, claim-level instructions/shots focused on the hard indicators above, and human-like calibration examples—not clause chopping or detect-recall patches that trade misses for false alarms. Given $N=86$ responses, small mean arm deltas should be interpreted cautiously (see power note above).

At a higher level, our methodology represents a highly promising first step toward scaling automated writing evaluations that are explainable, cost-effective, and accurate. Practically, by pairing binary presence scores with verbatim evidence spans, our approach creates an instant audit trail for automated grading that allows classroom teachers to verify AI-generated feedback at a glance, dramatically lowering the barrier to trust and deployment in live classrooms. Furthermore, our finding that joint prompting outperforms looped architectures translates to an immediate operational benefit, as developers can obtain multidimensional feedback in a single API call rather than N individual queries, reducing processing costs and network latency, and

making real-time, student-facing formative scoring economically and technically viable at scale. Specifically, to optimize both detection stability and extraction precision, we suggest a hybrid, two-step pipeline for operational designs: first, run a one-phase prompt for detection, ignoring the initial spans generated; second, run the second phase of the two-phase prompt to extract evidence for the indicators detected in the first step. This avoids wasting token overhead or model attention on absent concepts, focusing the computational budget entirely on retrieving high-quality, verbatim evidence.

A key theoretical implication of our work is that LLMs cannot simply bypass the deep pedagogical preparation traditionally required for human raters. In high-stakes educational testing, human scorers undergo rigorous training and rely on pages of scoring materials, anchor charts, and concrete examples illustrating when to apply specific codes. We observed that LLMs face a parallel bottleneck: while they easily classify the presence of a rubric concept, they fail to anchor the exact verbatim evidence without this same expert guidance and exemplary context. Consequently, the path to robust automated scoring is not through structural pipeline engineering, but through translating human calibration materials into model instructions. A crucial next step for our research will consist of systematically developing these detailed scoring annotations and empirically comparing LLM performance with and without these comprehensive calibration materials.

Limitations

This study has several limitations. First, the analysis sample is modest ($N=86$ responses; 946 indicator cells that are clustered within responses), so small mean differences between arms are often underpowered (Section 4.6); larger multi-task samples would be needed to confirm architecture effects. Second, we evaluate a single ONPAR earth-science prompt and one final model family (Gemini 3.1-pro few-shot); results may not transfer to other domains, grade bands, or open-weight models (Llama was dropped after hallucination issues on this task). Third, the oracle is a two-rater consensus with spans that themselves encode a particular evidentiary grain; disagreement with the model is not always “model error” versus an equally defensible alternate span. Fourth, Fuzzy-76 and Cont&76%

Rate are operational metric choices: lowering the similarity cutoff or accepting containment raises apparent extract agreement without changing the underlying quotes. Fifth, rubric C annotations and paraphrased shots were informed by error analysis on related runs; while shots paraphrase student language, residual optimism about annotation benefits is possible and should be tested on a fully held-out task. Sixth, we do not evaluate downstream feedback quality with teachers or students—evidence agreement is necessary but not sufficient for formative usefulness. Seventh, we do not report human–human span agreement on the same Fuzzy-76 / Cont&76% metrics, which would provide a ceiling for interpretability. Finally, these samples were collected in 2019 and there have been significant changes in student classroom writing practices since that time. The use of generative AI, in particular, may affect how well the findings apply to current formative assessment.

References

- Max Bachmann. 2025. [rapidfuzz/rapidfuzz: Release 3.13.0](#).
- Kadriye Ercikan and Daniel F. McCaffrey. 2022. [Optimizing implementation of artificial-intelligence-based automated scoring: An evidence-centered design approach for designing assessments for AI-based scoring](#). *Journal of Educational Measurement*, 59(3):272–287.
- Filip J. Kucia, Anirban Chakraborty, and Anna Wróblewska. 2026. [LLM essay scoring under holistic and analytic rubrics: Prompt effects and bias](#). *arXiv preprint arXiv:2604.00259*.
- Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ning Liu, and Xiaoming Zhai. 2024. [Applying large language models and chain-of-thought for automatic scoring](#). *Computers and Education: Artificial Intelligence*, 6:100213.
- Alessia Marigo, Laura Wright, and Linda Malkin. 2026. Scoring student science responses with small training datasets: Comparing LLMs and fine-tuned models. Poster presented at the annual meeting of the National Council on Measurement in Education (NCME). Poster presentation.
- NGSS Lead States. 2013. [Next Generation Science Standards: For States, By States](#). The National Academies Press, Washington, DC.
- Christopher Ormerod. 2025. [Automated essay scoring incorporating annotations from automated feedback systems](#). In *Proceedings of the Artificial Intelligence*

in Measurement and Education Conference (AIME-Con): Coordinated Session Papers, pages 9–18. National Council on Measurement in Education.

Yucheng Wang, Zhiyao Ding, Xuansheng Wu, Shengwei Sun, Ning Liu, and Xiaoming Zhai. 2026. [AutoSCORE: Enhancing automated scoring with multi-agent large language models via structured component recognition](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 40898–40906.

Benedikt Wisniewski, Klaus Zierer, and John Hattie. 2020. [The power of feedback revisited: A meta-analysis of educational feedback research](#). *Frontiers in Psychology*, 10:3087.

Laura Wright, H. Harkins, Rebecca Kopriva, W. Auty, Linda Malkin, and B. Myers. 2022. [Technology-enhanced tasks to assess three-dimensional science sense-making: Possibilities and lessons learned from the ONPAR NGSS-based classroom assessment project](#). *Contemporary Issues in Technology and Teacher Education*, 22(2).

Mingfeng Xue, X. Xiao, Y. Liu, and Mark Wilson. 2026. [On the consistency of automatic scoring with large language models](#). *Educational and Psychological Measurement*. Article 00131644261418138.

Fatih Yavuz, Özkan Çelik, and Gülşah Yavaş Çelik. 2025. [Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments](#). *British Journal of Educational Technology*, 56(1):150–166.

Junyi Zhou, Qiyuan Zhang, Yufei Wang, Fuyuan Lyu, Yidong Ming, Can Xu, Qingfeng Sun, Kai Zheng, Peng Kang, Xue Liu, and Chen Ma. 2026. [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)*, pages 31179–31200, San Diego, California, United States. Association for Computational Linguistics.