informatics

MDPI

Article

RubricAdapt-LLM: Measuring Criterion-Level Adaptation and Score Shifts Under Alternative Grading Rubrics

Catalin Anghel, Andreea Alexandra Anghel, Adina Cocu *, Marian Viorel Craciun, Simona Moldovanu, Mihai Vlase and Cristian Sandu

Department of Computer Science and Information Technology, "Dunărea de Jos" University of Galati, Științei St. 2, 800146 Galati, Romania; catalin.anghel@ugal.ro (C.A.); aa481@student.ugal.ro (A.A.A.); marian.craciun@ugal.ro (M.V.C.); simona.moldovanu@ugal.ro (S.M.); mihai.vlase@ugal.ro (M.V.); cristian.sandu@ugal.ro (C.S.)

* Correspondence: adina.cocu@ugal.ro

Abstract

Background: Large language models are increasingly used for rubric-based grading, but it remains unclear whether they adapt selectively when criterion weights change while the response, criterion definitions, and total score remain fixed. This study examined whether alternative point allocations produce targeted criterion-level adaptation or broader grading instability. **Methods:** A controlled paired design was applied to 1000 student responses, comprising 500 technical and 500 argumentative answers. Eight local open-weight LLMs evaluated every response under two analytic rubrics totaling 10 points. One point was transferred from Clarity to Completeness for technical responses and from Clarity to Dialecticality for argumentative responses. Of 16,000 expected evaluations, 15,999 were structurally valid, yielding 7999 complete cross-rubric pairs. Model outputs were analyzed for total-score shifts, affected-criterion adaptation, stability of unaffected criteria, model–human alignment, and correspondence with differences between two human evaluation conditions. **Results:** All models assigned lower mean scores under Rubric B, with mean shifts ranging from −1.239 to −0.182 points. Adaptation mechanisms differed substantially across models and response types. gemma3:4b frequently preserved technical total scores through compensating criterion changes, whereas llama3.1:8b showed extensive spillover into unaffected criteria. The Qwen models generally produced smaller total-score reductions and greater stability in unchanged dimensions. The mean score was 0.637 points higher under the human Rubric B condition than under the human Rubric A condition, although the two conditions were applied by different evaluator pairs; every model shifted negatively, and the lowest overall shift error was obtained by qwen3:4b at 1.498 points. **Conclusions:** Rubric sensitivity did not consistently imply localized or criterion-consistent adaptation. Reliable evaluation of LLM graders therefore requires separate analysis of total scores, affected criteria, unaffected criteria, human alignment, and cross-condition shift correspondence.

Keywords: large language models; automated grading; analytic rubrics; criterion-level adaptation; rubric sensitivity; score stability; human–AI alignment; open-weight models

Check for updates

Academic Editor: Antony Bryant

Received: 28 July 2026
Revised: 5 September 2026
Accepted: 9 September 2026
Published: 10 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Large language models have been evaluated for criterion-based grading, in which student performance is judged against predefined assessment criteria [1]. In analytic

assessment, a rubric separates performance into distinct criteria and specifies performance levels for each criterion [2]. When alternative rubric conditions preserve the same overall scale but redistribute points across criteria, a reliable grader should adjust the contribution of the affected criteria without reinterpreting unrelated dimensions of the same answer. This requirement is non-trivial, because LLM-based evaluators have been found to confuse distinct quality criteria in NLG evaluation settings [3]. RubricAdapt-LLM investigates whether alternative rubric weighting produces targeted, criterion-consistent score shifts or broader grading instability.

1.1. Background and Motivation

Automated assessment of constructed responses draws on natural language processing, machine learning, and deep learning approaches [4]. Research encompasses holistic, prompt-specific, cross-prompt, and trait-specific assessment, together with automated feedback generation [5]. LLM-based applications span essay and short-answer assessment [6], including medical and software engineering education [7,8], and bioinformatics [9]. These applications motivate systems capable of evaluating natural-language answers that may express relevant knowledge through different formulations.

LLMs have been investigated for formative assessment and feedback generation [10,11]. Research on interpretable short-answer scoring also examines how model evidence could support collaboration between automated and human graders [12]. Nevertheless, promising grading performance does not imply uniform reliability across evaluation settings [13,14]. Agreement, consistency, severity, fairness, and criterion fidelity must therefore be treated as related but distinct properties of an automated grader.

Rubrics also have educational and measurement functions that extend beyond converting judgments into numerical scores. A meta-analysis found a moderate positive effect of rubric use on academic performance, together with smaller pooled effects on self-regulated learning and self-efficacy [15]. Research on rubric validity emphasizes the importance of defining observable evidence for different levels of response complexity and aligning the rubric with the intended assessment objectives [16]. In automated text evaluation, LLM-Rubric demonstrated how separate rubric questions can operationalize multiple quality dimensions, while also showing that uncalibrated LLM predictions may disagree with human judgments [17]. Consequently, the presence of a rubric in the grading prompt does not by itself demonstrate that each component of that rubric governs the resulting evaluation as intended.

Existing rubric-based studies have examined the reliability and validity of LLM scores under a fixed analytic rubric, including assessments of EFL essays [18]. Criterion-specific investigations have also found that LLM evaluations can remain consistent for certain dimensions while exhibiting substantially lower agreement for others, with results affected by both prompt strategy and model architecture [19]. These findings leave unresolved whether an LLM grader responds locally and proportionally when the scoring importance of one criterion changes while the student answer and the remaining criteria are held constant. Distinguishing such legitimate rubric sensitivity from unintended changes in otherwise unaffected judgments provides the central motivation for RubricAdapt-LLM.

1.2. Related Work and Research Gap

Recent research has increasingly operationalized LLM-based grading through explicit criteria rather than a single holistic instruction. Across educational and broader evaluation settings, recent approaches have used criterion-level rubrics, multidimensional rubric-based metrics, question-specific scoring and deduction rules, and customized fine-grained evaluation criteria [20–23]. These studies demonstrate that rubric content can

structure an LLM evaluation, but they primarily assess performance under a selected or generated rubric. They do not test whether the same evaluator changes only the intended criterion contributions when two otherwise equivalent rubrics assign different weights to those criteria.

Criterion-level grading also creates requirements that are not captured by total-score agreement. Recent multi-trait scoring approaches have combined trait-specific rubric information, LLM-generated rationales, teacher-aligned multidimensional evaluation, and explicit modeling of dependencies among trait representations [24–26]. This literature recognizes both heterogeneity and dependence among evaluation dimensions. However, neither trait-specific reasoning nor explicit dependency modeling determines whether judgments for formally unchanged criteria remain invariant when the maximum score of another criterion is altered.

Prompt-design studies provide further evidence that LLM scoring is conditional on how assessment instructions are expressed. Recent work has varied prompting strategies, reflective prompt refinement, automatic prompt engineering, and iterative rubric revision, showing that changes in examples, reasoning instructions, item context, and rubric formulation can alter scoring behavior [27–30]. These findings establish sensitivity to prompt construction and rubric formulation, but they combine changes in wording, examples, reasoning scaffolds, or rubric content. Such interventions are therefore different from a controlled change in criterion weight under an otherwise frozen evaluation prompt.

Reliability-oriented research has mainly examined repeated-run stability, agreement across models, and alignment with human graders. Recent studies have combined repeated LLM grading, intra- and inter-model consistency analysis, and comparisons of grading behavior across model architectures and human evaluators [31–33]. These results show that a grader may be reproducible without being interchangeable with other graders or fully aligned with the intended assessment standard. Repeated-score stability is therefore insufficient for establishing criterion fidelity under alternative rubric conditions.

The internal use of evaluation dimensions is especially relevant when rubric weights are modified. Recent studies have identified differences in implicit feature weighting, criteria-entanglement effects, and broader cognitive biases in LLM-based evaluation, indicating that nominally distinct assessment dimensions may not be processed independently [34–36]. These findings raise the possibility that a localized rubric change may alter the model's interpretation of additional dimensions rather than only the criterion whose scoring range has changed.

Taken together, prior work has examined rubric-guided grading, question-specific rubric generation, multidimensional scoring, prompt and rubric refinement, repeated-run consistency, model–human agreement, implicit feature weighting, and dependence among evaluation criteria. To our knowledge, it has not isolated a paired condition in which the question, student answer, criterion definitions, total score scale, and weights of all unrelated criteria remain constant while points are transferred between two specified criteria. This distinction is necessary to separate responsiveness to alternative criterion weights from broader variation in evaluator behavior. An expected total-score shift does not by itself demonstrate correct adaptation if criteria unaffected by the modification also change; conversely, an unchanged total score can conceal compensating shifts across criteria. The unresolved problem therefore concerns two simultaneous properties: responsiveness of the affected criteria to the revised weights and stability of the unaffected criteria under the alternative rubric.

1.3. Aim, Contributions, and Research Questions

The aim of RubricAdapt-LLM is to examine whether LLM-based graders adapt their evaluations in a targeted and criterion-consistent manner when criterion weights are modified under controlled paired conditions. For each model, the same question and student response are evaluated under two alternative analytic rubrics, both totaling 10 points. For technical responses, the maximum score assigned to clarity is reduced from three to two points, while the maximum score assigned to completeness is increased from two to three points. For argumentative responses, the same reduction is applied to clarity, while the maximum score assigned to dialecticality is increased from two to three points. All criterion definitions, the maximum scores of the remaining criteria, and all prompt components other than the modified point allocations are preserved. The model configuration, decoding parameters, response content, and inference procedure are also held constant. This design enables the observed changes to be examined specifically in relation to the controlled redistribution of criterion weights.

The study additionally evaluates condition-specific alignment with human assessments. Each rubric condition is associated with evaluations produced by the corresponding pair of human evaluators, allowing the scores generated by each model to be compared with human evaluations obtained under the same rubric. Because different evaluator pairs applied the two rubric conditions, differences between the human evaluation conditions may reflect both the alternative criterion weights and evaluator-specific scoring tendencies. The human evaluations are therefore used as rubric-matched comparison points, without assuming that all differences between the two human conditions are caused exclusively by the rubric modification. The study does not seek to determine which rubric represents a universally preferable grading standard.

The study makes four specific contributions. First, it applies a controlled paired design in which 1000 student responses are submitted to eight open-weight LLMs under two alternative rubric conditions, defining 16,000 expected model–response–rubric evaluation tasks. Second, it distinguishes adaptation in the criteria whose maximum scores are modified from variation in criteria whose definitions and point allocations remain unchanged. The analysis considers both total-score differences and the direction, magnitude, and localization of criterion-level changes, including cases in which compensating criterion shifts produce little or no change in the total score. Third, it evaluates each model against the human assessments corresponding to each rubric condition and examines the descriptive correspondence between LLM-generated score shifts and the differences observed between the two human evaluation conditions. Fourth, it compares cross-rubric adaptation, stability of unaffected criteria, and condition-specific human alignment across the eight evaluated models and between technical and argumentative responses. These contributions are limited to the paired experimental conditions and rubric modifications examined in this study.

The study addresses the following research questions:

RQ1. How closely does each LLM align with the rubric-matched human evaluations under each rubric condition?

RQ2. How do LLM-generated total scores change when one point is redistributed between two criteria while the total rubric score remains constant?

RQ3. What are the direction, magnitude, and localization of score changes in the criteria whose maximum scores are modified?

RQ4. To what extent do criteria whose definitions and maximum scores remain unchanged preserve stable scores across the two rubric conditions?

RQ5. How do cross-rubric score shifts, criterion-level stability, and their correspondence with differences between the human evaluation conditions vary across models and between technical and argumentative responses?

Together, these research questions assess whether alternative criterion weights produce targeted adaptation without destabilizing unchanged grading dimensions and whether the resulting score patterns remain aligned with the corresponding human evaluation conditions.

2. Materials and Methods

This section describes the controlled paired study design, the educational assessment dataset, the alternative rubric conditions and corresponding human evaluations, the evaluated models and inference protocol, the validation of the structured grading outputs, and the statistical procedures used to examine cross-rubric score changes.

2.1. Study Design

RubricAdapt-LLM was designed as a controlled paired secondary methodological analysis of 1000 short open-ended student responses. The assessment set contains 500 technical and 500 argumentative responses, corresponding to 100 students answering ten questions. Each response was submitted to eight open-weight LLMs under two alternative analytic rubric conditions, defining 16,000 model–response–rubric evaluation tasks.

The two rubric conditions originated from the scoring priorities established by two pairs of human evaluators for the same assessment set. Evaluators E1 and E2 applied Rubric A, whereas evaluators E3 and E4 applied Rubric B. Thus, the total score and criterion definitions remained unchanged, while one point was redistributed between two predefined criteria.

Each of the eight LLMs evaluated every response under both rubric conditions. The primary paired unit consisted of the two evaluations generated by the same model for the same student response under Rubric A and Rubric B. This design defined 8000 expected paired model–response comparisons before output validation. The human evaluations were used as condition-specific comparison points for assessing model alignment under each rubric.

Figure 1 summarizes the controlled paired workflow used to generate and organize the analytical data.

The structured outputs are subsequently validated and paired by student response, question, model, and rubric condition. The resulting dataset supports the analysis of total-score shifts, adaptation in the affected criteria, stability of the unaffected criteria, alignment with rubric-matched human evaluations, and differences across models and response types.

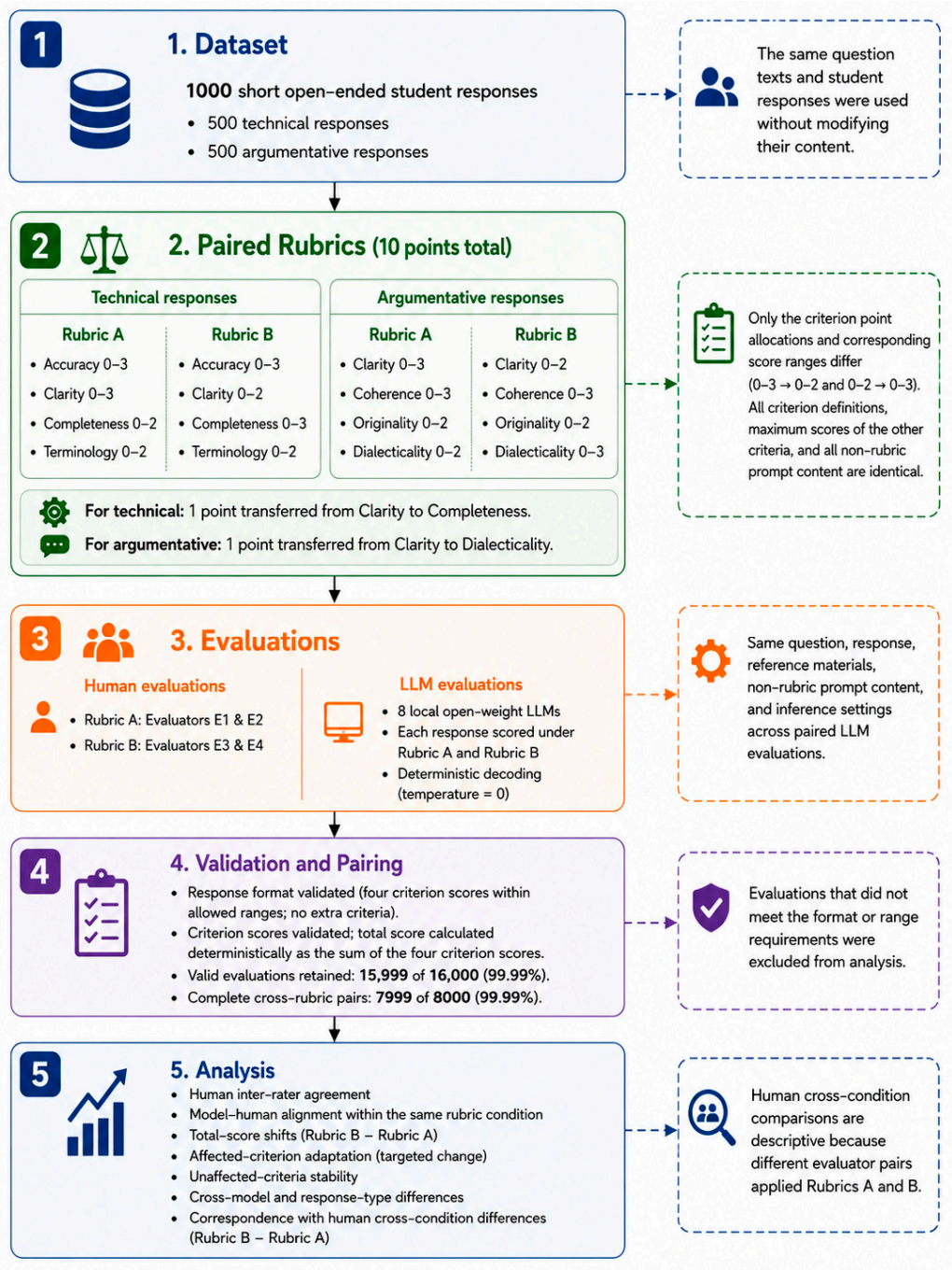


Figure 1. Controlled paired evaluation workflow used in RubricAdapt-LLM.

2.2. Dataset and Assessment Types

The educational assessment dataset was originally introduced in the GradeAgentOps study [37]. It contains 1000 short open-ended responses produced by 100 students across ten university-level assessment questions. Each student answered all ten questions, yielding 100 responses for each item. RubricAdapt-LLM uses the same question texts and student responses, without modifying their content.

Each record is identified by the corresponding student and question identifiers and contains the question text and the student's free-form response. These identifiers were preserved throughout data preparation and were used to align the two human evaluation conditions and the LLM-generated outputs at the student–question level.

The assessment set is divided equally into five technical and five argumentative questions. The technical questions address version control, artificial intelligence, ma-

chine learning, expert systems, and facial-recognition technologies. The argumentative questions require students to formulate and justify positions on ethical, educational, and socio-technical issues related to artificial intelligence and digital technologies. The resulting dataset therefore contains 500 technical and 500 argumentative responses. Table 1 summarizes the assessment questions and their classification.

Table 1. Assessment questions and response types.

Question	Assessment Type	Question Focus
Q1	Technical	Version control systems
Q2	Technical	Narrow artificial intelligence versus general artificial intelligence
Q3	Technical	Naive Bayes classifier
Q4	Technical	Search in expert systems
Q5	Technical	Facial recognition in security
Q6	Argumentative	Facial recognition in public spaces
Q7	Argumentative	GitHub versus Moodle for assignment management
Q8	Argumentative	Artificial intelligence and ethical decision-making
Q9	Argumentative	Personalization algorithms and individual autonomy
Q10	Argumentative	The Turing Test and machine intelligence

The two assessment types use different criterion sets. Technical responses are evaluated according to accuracy, clarity, completeness, and terminology, whereas argumentative responses are evaluated according to clarity, coherence, originality, and dialecticality. The same criterion definitions are used in both rubric conditions; the conditions differ only in the point allocations assigned to clarity and completeness for technical responses and to clarity and dialecticality for argumentative responses.

The balanced structure provides equal sample sizes for the two response types and allows cross-rubric score changes to be examined separately for technical and argumentative assessment tasks.

2.3. Alternative Rubric Conditions and Human Evaluations

The two analytic rubric conditions were defined by two pairs of human evaluators for the same assessment set. Evaluators E1 and E2 established Rubric A, whereas evaluators E3 and E4 established Rubric B. Each pair retained the same four assessment criteria for technical and argumentative responses and the same total score of 10 points, but assigned different relative importance to selected criteria when defining its scoring approach. The resulting rubric conditions therefore represent two human-defined weighting schemes applied to the same questions and student responses. Table 2 presents the complete criterion structure and point allocation used under each rubric condition.

Table 2. Criterion point allocations under Rubric A and Rubric B.

Response Type	Criterion	Rubric A	Rubric B
Technical	Accuracy	3	3
Technical	Clarity	3	2
Technical	Completeness	2	3
Technical	Terminology	2	2
Argumentative	Clarity	3	2

Table 2. *Cont.*

Response Type	Criterion	Rubric A	Rubric B
Argumentative	Coherence	3	3
Argumentative	Originality	2	2
Argumentative	Dialecticality	2	3

For technical responses, Rubric A allocated three points to accuracy, three points to clarity, two points to completeness, and two points to terminology. Under Rubric B, the allocations for accuracy and terminology remained unchanged, while clarity was reduced from three to two points and completeness was increased from two to three points. For argumentative responses, Rubric A allocated three points to clarity, three points to coherence, two points to originality, and two points to dialecticality. Rubric B preserved the allocations for coherence and originality, while clarity was reduced from three to two points and dialecticality was increased from two to three points. In both response types, the criterion definitions and total available score remained unchanged.

Human scoring was performed by four domain experts, all university academics in computer science with doctoral-level qualifications and experience in teaching and grading university-level computer science subjects. The evaluators had previously assessed the source educational corpus and were therefore familiar with its technical and argumentative tasks, the grading instructions, and the common 0–10 scoring scale. Before the individual grading process began, E1 and E2 agreed on how Rubric A would be applied, while E3 and E4 agreed on the application of Rubric B. This pair-level calibration concerned the scoring approach associated with the corresponding criterion weights. After this stage, each evaluator independently assessed the complete set of 1000 responses. Evaluators did not have access to the scores assigned by the other member of the same pair or by the evaluators working under the alternative rubric condition.

For each response, the evaluators assigned one score to each of the four criteria defined for the corresponding assessment type. The final score was calculated as the sum of the criterion-level scores and therefore remained on the common 0–10 scale. The human grading process did not use the reference answers, gold points, or banned misconceptions available in the source dataset. Human scores were based exclusively on the question, the student response, the applicable criterion definitions, and the point allocations specified by the corresponding rubric.

The four human evaluation sets were aligned at the student–question level. Scores from E1 and E2 provide the human comparison points for Rubric A, whereas scores from E3 and E4 provide the corresponding comparison points for Rubric B. These data are used to evaluate model–human alignment within each rubric condition and to examine whether score shifts produced by the LLMs correspond descriptively to the differences observed between the two human evaluation conditions.

Because the two rubric conditions were applied by different evaluator pairs, cross-condition differences in human scores may reflect both the alternative criterion weights and evaluator-specific scoring tendencies. The human comparison therefore supports condition-specific alignment and descriptive shift analysis, but does not permit all differences between Rubric A and Rubric B to be attributed exclusively to the redistribution of points.

2.4. Evaluated Models and Inference Protocol

The evaluated model set comprised eight open-weight LLMs from four model families, with two parameter scales represented within each family: Qwen3 [38,39], Gemma 3 [40,41], Llama 3.1 [42,43], and Phi-4 [44,45]. Table 3 lists the exact model identifiers used throughout the experiment.

Table 3. Open-weight models evaluated in RubricAdapt-LLM.

Model Family	Ollama Model Identifier	Nominal Parameter Scale
Qwen3	qwen3:4b	4B
Qwen3	qwen3:14b	14B
Gemma 3	gemma3:4b	4B
Gemma 3	gemma3:27b	27B
Llama 3.1	llama3.1:8b	8B
Llama 3.1	llama3.1:70b	70B
Phi-4	phi4-mini:3.8b	3.8B
Phi-4	phi4:14b	14B

All models were served locally through Ollama v0.32.1 [46] and were used without additional training, fine-tuning, or adaptation on the assessment dataset. Each combination of model, student response, and rubric condition was processed as an independent non-streaming generation call without conversational history. For every model–response pair, the evaluation under Rubric A and the evaluation under Rubric B were therefore generated separately.

The analytic grading prompt contained the assessment type, question text, student response, reference answer, expected content elements, and any banned misconceptions defined for the corresponding question. Expected content elements and their relative content weights were supplied as common evidence for interpreting the substantive content of the response and were not added as a separate numerical score. For argumentative questions, the models were explicitly instructed to evaluate reasoning quality without penalizing a defensible position solely because it differed from the conclusion presented in the reference answer. The reference materials were identical under both rubric conditions.

The prompt additionally specified the four criterion definitions, criterion-specific scoring guidance, and ordinal score anchors for criteria with maximum values of two or three points. Models were instructed to score the criteria independently, award partial credit when relevant evidence was present, and avoid transferring unused points between criteria. No human scores, evaluator identities, or committee labels were supplied to the models. Between the two paired rubric calls, only the criterion point allocations and the corresponding permissible score ranges were changed; the remaining prompt content was preserved.

Each model was required to return a single structured JSON object containing the four integer criterion scores, an auxiliary confidence category, and a brief rationale. The permissible range of each criterion score was constrained by the active rubric through a rubric-specific JSON schema. The model was not asked to generate the final grade. Instead, the final score was calculated deterministically as the sum of the four validated criterion scores, producing a value on the common 0–10 scale.

Inference was performed with temperature 0.0, top-p 1.0, top-k 40, seed 42, a context window of 16,384 tokens, and a maximum generation length of 2048 tokens. Explicit reasoning mode was disabled. The request timeout was set to 3600 s, and models were retained in memory for 30 min between calls. The same generation settings were used under both rubric conditions and across all responses.

Same-rubric repeatability was assessed by independently repeating the evaluation of all 1000 responses once for each of the eight models under each rubric. The original evaluations were denoted A1 and B1, and the repeated evaluations A2 and B2, yielding 16,000 additional inference calls. Each repeated evaluation used the same prompt text, rubric-specific output schema, and generation parameters as its corresponding original evaluation.

The experiments were executed on a 64-bit Windows 11 Education virtual machine using Python 3.12.10 and Ollama v0.32.1. The system was configured with an AMD EPYC 9654 processor (Advanced Micro Devices, Inc., Santa Clara, CA, USA) using 80 processing cores, 128 GB of RAM, and an NVIDIA L40S-48Q GPU with 48 GB of video memory (NVIDIA Corporation, Santa Clara, CA, USA). The execution environment used NVIDIA driver version 581.42 and CUDA 13.0.

2.5. Output Validation and Reproducibility

Each inference call was constrained by a rubric-specific JSON schema derived from the response type and the maximum score assigned to each criterion under the active rubric. The required output contained the criterion scores, an auxiliary confidence category, and a rationale. The confidence category was retained only as metadata and was not used in the calculation of the final score or in any statistical analysis. The scores object was required to contain exactly the four criteria applicable to the technical or argumentative response, without missing or additional fields. Criterion scores were restricted to integers between zero and the corresponding rubric-specific maximum. The confidence field accepted only low, medium, or high, while the rationale was required to contain a non-empty textual explanation.

Generated outputs were processed through deterministic structural parsing and validation. Direct JSON responses were parsed without transformation. A limited recovery procedure was applied when the same JSON object was enclosed in a Markdown code block or accompanied by additional text. This procedure modified only the response envelope needed for JSON extraction and did not alter criterion scores, confidence values, or rationale content. Outputs that remained incompatible with the required structure, contained incorrect criterion names, used non-integer scores, exceeded the permitted score ranges, or omitted required content were classified as parse errors.

The final score was not accepted as a model-generated field. For every structurally valid output, it was calculated deterministically as the sum of the four validated criterion scores. This procedure ensured that the final value remained on the 0–10 scale and was internally consistent with the criterion-level evaluation. Valid records were assigned the status ok, while unsuccessful model requests and structurally invalid outputs were recorded separately as request_error and parse_error, respectively.

Ollama request failures were retried up to two times after the initial attempt, with a fixed five-second interval between attempts. Structural parse errors were not semantically corrected or converted into valid evaluations. Only successfully validated records were marked as completed. When an interrupted or incomplete run was resumed, completed tasks were skipped, whereas tasks without a valid record remained eligible for re-execution.

Each evaluation task was identified by the combination of model, response identifier, and rubric condition. The saved record also contained the student and question identifiers, assessment type, criterion maxima, criterion scores, calculated final score, confidence category, rationale, validation status, parsing method, error information, number of request attempts, processing time, and creation timestamp. For traceability, the complete prompt, raw model response, selected reference materials, and Ollama generation metadata were retained with each record.

Run-level artifacts were stored separately using timestamped identifiers. These artifacts comprised the line-delimited JSON inference output, the resolved experimental configuration, the execution log, and summary statistics for the corresponding invocation. The configuration record preserved the selected models, rubric definitions, criterion definitions, scoring guidance, score-level anchors, inference parameters, dataset path, and expected number of tasks. Before execution, the selected models were checked against

the locally available Ollama models. Resume operations were rejected when the existing output contained records generated under a different prompt version, preventing results from different experimental specifications from being combined.

Before statistical analysis, the final output dataset was audited against the 16,000 expected combinations of 1000 responses, eight models, and two rubric conditions. The integrity checks verified task uniqueness, the expected number of records for each model and rubric, the correspondence between criterion names and assessment type, compliance with rubric-specific score ranges, and equality between each final score and the sum of its criterion scores. Only structurally valid records were eligible for statistical analysis. Cross-rubric analyses additionally required valid outputs under both rubric conditions for the same model–response pair, and missing or invalid scores were not imputed.

2.6. Rubric-Adaptation Metrics and Statistical Analysis

The statistical analysis was conducted on paired evaluations generated for the same student response by the same model under Rubric A and Rubric B. Let s_{imrc} denote the score assigned to criterion c for response i, model m, and rubric condition $r \in \{A, B\}$, and let M_{rc} denote the maximum score permitted for that criterion under the corresponding rubric. Here, i indexes the 1000 student responses, m indexes the eight evaluated models, r identifies the rubric condition, and c identifies one of the four criteria applicable to the response type.

The total score assigned by model m to response i under rubric condition r was calculated as the sum of the four criterion scores:

$$T_{imr} = \sum_{c=1}^{4} s_{imrc}. \quad (1)$$

Thus, T_{imA} and T_{imB} represent the total scores assigned to the same response by the same model under Rubric A and Rubric B, respectively. Because both rubrics total 10 points, T_{imr} remains on the common 0–10 scale under both conditions.

The cross-rubric total-score shift was defined as:

$$\Delta T_{im} = T_{imB} - T_{imA} \quad (2)$$

For a given response i and model m, a positive ΔT_{im} indicates that the model assigned a higher total score under Rubric B, a negative value indicates a lower score, and a value of zero indicates that the final grade remained unchanged. The distribution of ΔT was summarized using the mean, median, interquartile range, and the proportions of positive, negative, and zero shifts. These indicators were calculated overall, separately for each model, and independently for technical and argumentative responses.

Criterion-level changes were first examined using the raw score difference:

$$\Delta s_{imc} = s_{imBc} - s_{imAc} \quad (3)$$

Here, Δs_{imc} represents the change in the number of points awarded to criterion c when the same response was evaluated under Rubric B instead of Rubric A. Raw criterion changes show how each dimension contributed to the observed change in the final 10-point score.

Because the maximum score of the affected criteria changes between the two rubric conditions, raw scores alone are not sufficient for comparing the model's relative assessment

of criterion quality. Each criterion score was therefore also normalized by the maximum score available under the corresponding rubric:

$$p_{imrc} = \frac{s_{imrc}}{M_{rc}}, 0 \leq p_{imrc} \leq 1 \tag{4}$$

The quantity p_{imrc} represents the proportion of the available score awarded to criterion c. A value of 0 indicates that no credit was awarded, whereas a value of 1 indicates that the model awarded the maximum score available for that criterion under the active rubric. The normalized criterion-level shift was calculated as:

$$\Delta p_{imc} = p_{imBc} - p_{imAc} \tag{5}$$

Because Δp_{imc} is a signed difference between two normalized scores, its range is $[-1, 1]$. Negative values indicate a lower normalized criterion score under Rubric B, positive values indicate a higher normalized criterion score, and zero indicates no change.

For example, a raw score of 2 for a criterion with a maximum of 3 corresponds to a normalized score of $2/3$, whereas the same raw score under a maximum of 2 corresponds to a normalized score of 1. Reporting both Δs_{imc} and Δp_{imc} therefore distinguishes changes in the effective point contribution from changes in the proportion of the available criterion score awarded by the model.

The affected and unaffected criterion sets depended on the response type. For technical responses, clarity and completeness formed the affected set, whereas accuracy and terminology formed the unaffected set. For argumentative responses, clarity and dialecticality were affected, whereas coherence and originality were unaffected. These sets are denoted by $\mathcal{C}_i^{\text{aff}}$ and $\mathcal{C}_i^{\text{unaff}}$, respectively, where the applicable set is determined by the type of response i.

To separate intended adaptation from broader grading variation, profile distance was also calculated independently for the affected and unaffected criteria:

$$D_{im}^{\text{aff}} = \frac{1}{2} \sum_{c \in \mathcal{C}_i^{\text{aff}}} \mid p_{imBc} - p_{imAc} \mid, \tag{6}$$

$$D_{im}^{\text{unaff}} = \frac{1}{2} \sum_{c \in \mathcal{C}_i^{\text{unaff}}} \mid p_{imBc} - p_{imAc} \mid \tag{7}$$

Thus, D_{im}^{aff} measures the mean normalized change in the two criteria whose point allocations were modified, whereas D_{im}^{unaff} measures the corresponding change in the two criteria whose definitions and maximum scores remained identical. Both distances range from 0 to 1.

Stability of the unaffected criteria was evaluated at two levels. For an individual unaffected criterion c, exact stability was recorded when:

$$\Delta s_{imc} = 0. \tag{8}$$

Joint unaffected-criterion stability was recorded when both unaffected criterion scores remained identical across Rubric A and Rubric B. The joint stability rate was defined as the proportion of paired evaluations satisfying this condition. Its complement was defined as the spillover rate and represents the proportion of cases in which at least one unaffected criterion changed despite having the same definition and maximum score under both rubric conditions.

Directional adaptation was examined using the raw criterion changes because raw scores determine the final grade. For technical responses, an evaluation was classified as directionally concordant when clarity did not increase, completeness did not decrease, and at least one of the two affected scores changed. Formally, this condition required:

$$\Delta s_{im,\text{clarity}} \leq 0, \Delta s_{im,\text{completeness}} \geq 0, \tag{9}$$

with at least one strict inequality. For argumentative responses, the corresponding condition was:

$$\Delta s_{im,\text{clarity}} \leq 0, \Delta s_{im,\text{dialecticality}} \geq 0, \tag{10}$$

again with at least one affected criterion changing. Evaluations in which clarity increased or the criterion whose maximum was increased received a lower score were classified as directionally discordant. Cases in which neither affected criterion changed were classified separately as insensitive to the localized rubric modification. Directionally concordant cases were further divided into fully concordant cases, in which Clarity decreased and the criterion whose maximum increased also increased, and partially concordant cases, in which only one of these expected changes occurred while the other affected criterion remained unchanged.

Rubric insensitivity was evaluated at the full-profile level. A paired evaluation was classified as fully rubric-insensitive when all four raw criterion scores were identical under Rubric A and Rubric B:

$$s_{imAc} = s_{imBc} \text{ for all } c. \tag{11}$$

The rubric-insensitivity rate was calculated as the proportion of model–response pairs satisfying this condition. This indicator was interpreted descriptively because an unchanged profile may be compatible with the content of a particular response, especially when criterion scores are already at the floor, ceiling, or at values permitted under both rubrics.

A separate compensating-shift indicator was used to identify cases in which the final score remained unchanged even though the internal criterion profile changed. A compensating shift was recorded when:

$$\Delta T_{im} = 0 \tag{12}$$

and at least one $\Delta s_{imc} \neq 0$. The compensating-shift rate therefore measures the proportion of evaluations in which criterion-level changes would be concealed by analysis of the final score alone.

Human inter-rater agreement was examined separately within each rubric condition before model–human comparisons were performed. Under Rubric A, agreement between E1 and E2 was quantified using signed bias, mean absolute error (MAE), root mean squared error (RMSE), quadratic weighted kappa (QWK), and the two-way random-effects, absolute-agreement, single-measure intraclass correlation coefficient, ICC(2,1). The same indicators were calculated between E3 and E4 under Rubric B. Criterion-level human agreement was additionally summarized using exact agreement and MAE for each criterion. These analyses establish the amount of human variability present within each rubric condition and avoid treating either individual evaluator as an error-free ground truth.

Model–human alignment was calculated only between scores produced under the same rubric condition. Under Rubric A, each model was compared separately with E1 and E2. Under Rubric B, each model was compared separately with E3 and E4. Total-score alignment was quantified using signed bias, MAE, RMSE, QWK, and exact agreement.

The evaluator-specific results were retained separately to preserve the human variability observed within each rubric condition.

Same-rubric repeatability was quantified separately for each model using valid A1–A2 and B1–B2 pairs. Total-score comparisons used signed bias (repeat minus original), MAE, QWK, and exact agreement; full-profile repeatability required identical scores on all four criteria. Joint stability and normalized profile distance for the two criteria unaffected by rubric redistribution were also calculated within each rubric, separately for technical and argumentative responses. Comparisons with the corresponding A1–B1 measures used records valid in all four evaluations. These same-rubric measures provided an empirical baseline for interpreting cross-rubric variation, without isolating a causal effect of rubric redistribution.

To compare score shifts across the two human evaluation conditions, the mean human score was calculated separately for each response and rubric. Let H_{iA} denote the mean of the scores assigned by E1 and E2 to response i, and let H_{iB} denote the mean of the corresponding scores assigned by E3 and E4:

$$H_{iA} = \frac{E1_i + E2_i}{2}, \; H_{iB} = \frac{E3_i + E4_i}{2} \tag{13}$$

The human cross-condition score difference was then defined as:

$$\Delta H_i = H_{iB} - H_{iA} \tag{14}$$

Thus, ΔH_i represents the difference between the mean human score obtained under Rubric B and the mean human score obtained under Rubric A for the same response. It does not represent a causal rubric effect because the two conditions were applied by different evaluator pairs.

For each model m, correspondence between the model-generated shift and the human cross-condition difference was evaluated using the mean absolute shift error:

$$\mathrm{MAE}_{\Delta,m} = \frac{1}{N}\sum_{i=1}^{N} \mid \Delta T_{im} - \Delta H_i \mid, \tag{15}$$

where N is the number of responses with valid paired model outputs and complete human evaluations. Lower values indicate closer correspondence between the model-generated score shift and the human cross-condition difference in both direction and magnitude.

The mean model shift and the mean human cross-condition difference were reported descriptively, together with their signed difference. Exact shift agreement was defined as the proportion of responses for which the model-generated shift was identical to the corresponding human cross-condition difference:

$$\Delta T_{im} = \Delta H_i \tag{16}$$

Because model total scores were integer-valued, model-generated shifts also took integer values. In contrast, the human cross-condition difference was calculated from the means of two evaluators under each condition and could therefore take half-point values. Exact shift agreement was consequently attainable only when the human cross-condition difference was integer-valued and was interpreted as a strict descriptive measure rather than as a scale-independent agreement indicator.

Mean absolute shift error was summarized overall and separately for technical and argumentative responses, whereas exact shift agreement was reported overall. These

cross-condition comparisons were interpreted descriptively because differences between the human conditions may reflect both rubric weighting and evaluator-specific scoring tendencies.

All analyses were conducted on paired valid records. A model–response pair was included in cross-rubric analyses only when valid outputs were available under both Rubric A and Rubric B. Model–human analyses additionally required the relevant human evaluations to be present. Missing or invalid scores were not imputed. The number and distribution of excluded records, if any, were reported for each analysis.

Uncertainty was quantified using 10,000 student-cluster bootstrap resamples. In each resample, 100 students were sampled with replacement, and all ten responses and associated human and model evaluations for each selected student were retained together. This preserved the dependence among responses from the same student and the paired structure of the two rubric conditions. Percentile-based 95% confidence intervals were then derived from the resulting bootstrap distributions.

Model-level comparisons were based on student-level aggregate values so that each student, rather than each individual response, represented the independent analytical unit. This aggregation avoided treating the multiple responses contributed by the same student as independent observations. Overall differences among the eight models were examined using the Friedman test because the same students contributed paired observations for all models and the analysis did not require a normality assumption for the bounded outcome measures. Kendall's W was reported to quantify the magnitude of the overall cross-model effect. The six Friedman p-values corresponding to the three outcomes evaluated separately for technical and argumentative responses were adjusted jointly using the Holm procedure to control the family-wise error rate across the six related cross-model comparisons. When a Friedman comparison was statistically significant, paired Wilcoxon signed-rank tests were used for the 28 model-to-model comparisons because each contrast involved the same student-level units. Holm adjustment was applied separately within each of the six post hoc families. Technical–argumentative comparisons were also performed using paired student-level aggregates because each student contributed responses to both assessment types. The 24 within-model p-values corresponding to eight models and three outcomes were adjusted jointly using the Holm procedure. Two-sided adjusted p-values below 0.05 were considered statistically significant, while effect estimates and 95% confidence intervals were interpreted together with the significance tests. The analysis emphasized the magnitude and localization of cross-rubric changes rather than statistical significance alone.

3. Results

The analyses were conducted on 1000 student responses, comprising 500 technical and 500 argumentative answers. Each response was graded by eight open-weight LLMs under both rubric conditions, while four human evaluators provided two independent reference assessments for each condition. Of the 16,000 initial model evaluations, 15,999 passed the predefined structural validation procedure; one structurally invalid output was excluded. This resulted in 7999 complete model–response pairs for the analyses requiring direct comparison between Rubrics A and B, while condition-specific analyses retained all valid evaluations available under the corresponding rubric. The repeatability runs A2 and B2 yielded 16,000 structurally valid evaluations, providing 8000 valid A1–A2 pairs and 7999 valid B1–B2 pairs. Because each student contributed ten responses, the dependence among responses from the same student was preserved in the statistical analyses through student-level aggregation and cluster bootstrap resampling.

3.1. Human Inter-Rater Agreement Under the Two Rubric Conditions

Human scores cannot be treated as perfectly interchangeable reference labels when two evaluators applying the same rubric may differ in both overall scoring level and criterion-level point allocation. Inter-rater agreement therefore provides the reference context required to determine whether subsequent model–human discrepancies remain within, or extend beyond, the variability observed between human evaluators.

Agreement was examined separately for the two rubric conditions. E1 and E2 independently evaluated all 1000 student responses under Rubric A, whereas E3 and E4 evaluated the same responses under Rubric B. The analysis was performed both overall and separately for the 500 technical and 500 argumentative responses. Signed bias was calculated as the score assigned by the second evaluator minus the score assigned by the first evaluator. Positive values therefore indicate higher scores from E2 under Rubric A and from E4 under Rubric B. Table 4 reports the evaluator means together with the error-based and agreement measures calculated on the final 0–10 score.

Table 4. Human inter-rater agreement on the final 0–10 score.

Rubric and Evaluator Pair	Response Type	*n*	Mean Scores, First/Second	Bias [95% CI]	MAE/RMSE	QWK/ICC(2,1)	Exact (%)
A, E1–E2	Overall	1000	4.746/5.415	0.669 [0.507, 0.833]	1.827/2.296	0.678/0.678	16.7
A, E1–E2	Technical	500	5.190/6.038	0.848 [0.660, 1.038]	1.828/2.310	0.654/0.655	18.2
A, E1–E2	Argumentative	500	4.302/4.792	0.490 [0.276, 0.706]	1.826/2.283	0.678/0.678	15.2
B, E3–E4	Overall	1000	5.239/6.195	0.956 [0.861, 1.050]	1.304/1.825	0.783/0.783	28.3
B, E3–E4	Technical	500	5.844/6.892	1.048 [0.902, 1.194]	1.516/2.131	0.711/0.711	25.0
B, E3–E4	Argumentative	500	4.634/5.498	0.864 [0.748, 0.980]	1.092/1.456	0.840/0.841	31.6

Under Rubric A, E1 assigned an overall mean score of 4.746, whereas E2 assigned a mean score of 5.415. The resulting positive bias of 0.669 points indicates a systematic tendency for E2 to assign higher final scores. The bootstrap confidence interval remained above zero, showing that this directional difference was consistently reproduced across student-level resamples. At the individual-response level, however, the evaluator differences were not limited to a constant upward shift. The overall MAE was 1.827 points and the RMSE was 2.296 points, indicating that the scores frequently differed by more than one point and that some responses produced substantially larger discrepancies. QWK and ICC(2,1) were both 0.678, while exact agreement on the final score was obtained for 16.70% of the responses.

The separation by response type shows that the higher scoring tendency of E2 was more pronounced for technical responses. The technical bias was 0.848 points, compared with 0.490 points for argumentative responses. Nevertheless, the corresponding MAE values were almost identical, at 1.828 and 1.826 points, and the RMSE values also remained close. The principal difference between the two subsets therefore concerned the direction of the score difference rather than its average absolute magnitude. Agreement was slightly lower for technical responses, with QWK = 0.654 and ICC(2,1) = 0.655, compared with 0.678 for both measures in the argumentative subset.

Under Rubric B, E3 assigned an overall mean score of 5.239 and E4 assigned a mean score of 6.195. The positive bias of 0.956 points was larger than that observed within the Rubric A pair and remained positive for both response types. Despite this larger difference in mean scoring level, E3 and E4 produced smaller response-level errors, with an overall MAE of 1.304 points and RMSE of 1.825 points. Their QWK and ICC(2,1) values were both 0.783, and exact agreement reached 28.30%.

This combination of a larger bias and stronger agreement suggests that the difference between E3 and E4 contained a stronger systematic component than the difference observed

between E1 and E2. Bias measures the direction and average size of the difference, whereas MAE, RMSE, QWK, and ICC(2,1) describe how closely the paired scores correspond across individual responses. The E3–E4 pair therefore differed more clearly in overall scoring severity but preserved a closer response-level scoring pattern.

Agreement under Rubric B was stronger for argumentative than for technical responses. For the argumentative subset, the MAE decreased to 1.092 points and the RMSE to 1.456 points, while QWK and ICC(2,1) increased to 0.840 and 0.841, respectively. Exact agreement reached 31.60%. Technical responses produced a larger MAE of 1.516 points, a larger RMSE of 2.131 points, and QWK and ICC(2,1) values of 0.711. The positive bias remained substantial in both subsets, reaching 1.048 points for technical responses and 0.864 points for argumentative responses.

The values obtained for the two rubric conditions should be interpreted as characteristics of their respective evaluator pairs. Because Rubric A and Rubric B were applied by different evaluators, the differences between E1–E2 and E3–E4 cannot be attributed exclusively to the alternative criterion weights. They may also reflect evaluator-specific differences in scoring severity and in the interpretation of individual response characteristics.

Total-score agreement does not indicate how the observed differences were distributed across the four analytic criteria. Two evaluators may assign similar final scores while distributing points differently, or they may produce a positive total-score bias concentrated in only one or two dimensions. Criterion-level agreement was therefore examined separately for technical and argumentative responses. Table 5 reports the mean scores assigned by the two evaluators, the MAE, and the exact agreement rate for each active criterion.

Table 5. Criterion-level human inter-rater agreement under the two rubric conditions.

Rubric and Evaluator Pair	Response Type	Criterion (Maximum)	Mean Scores, First/Second	MAE	Exact (%)
A, E1–E2	Technical	Accuracy (3)	1.364/1.812	0.916	41
A, E1–E2	Technical	Clarity (3)	2.542/2.026	0.564	52.2
A, E1–E2	Technical	Completeness (2)	0.560/1.012	0.564	50.8
A, E1–E2	Technical	Terminology (2)	0.724/1.188	0.592	44.6
A, E1–E2	Argumentative	Clarity (3)	2.276/1.718	0.598	46
A, E1–E2	Argumentative	Coherence (3)	1.210/1.442	0.844	37.2
A, E1–E2	Argumentative	Originality (2)	0.300/0.798	0.558	55.4
A, E1–E2	Argumentative	Dialecticality (2)	0.516/0.834	0.518	55
B, E3–E4	Technical	Accuracy (3)	1.668/1.918	0.498	59.2
B, E3–E4	Technical	Clarity (2)	1.658/1.726	0.272	78.2
B, E3–E4	Technical	Completeness (3)	1.336/1.920	0.704	37.8
B, E3–E4	Technical	Terminology (2)	1.182/1.328	0.314	69.4
B, E3–E4	Argumentative	Clarity (2)	1.466/1.770	0.388	61.2
B, E3–E4	Argumentative	Coherence (3)	1.354/1.592	0.41	63
B, E3–E4	Argumentative	Originality (2)	0.558/1.144	0.586	48.8
B, E3–E4	Argumentative	Dialecticality (3)	1.256/0.992	0.508	57.4

Under Rubric A, the higher technical scores assigned by E2 were not produced by a uniform increase across all four criteria. E2 assigned higher mean scores for Accuracy, Completeness, and Terminology, but assigned a lower mean score for Clarity. Accuracy showed the largest technical discrepancy, with evaluator means of 1.364 and 1.812, an MAE of 0.916 points, and exact agreement of 41.00%. Clarity and Completeness produced the same MAE of 0.564 points, although their differences occurred in opposite directions. E1 assigned higher Clarity scores, whereas E2 assigned higher Completeness scores. The positive technical total-score bias therefore resulted from increases in three dimensions that were partially offset by lower Clarity scores.

The Rubric A argumentative results followed a similar criterion-specific pattern. E1 assigned a higher mean Clarity score than E2, at 2.276 compared with 1.718. In contrast, E2 assigned higher mean scores for Coherence, Originality, and Dialecticality. Coherence showed the largest disagreement in this subset, with an MAE of 0.844 points and the lowest exact agreement rate, at 37.20%. Originality and Dialecticality showed the highest exact agreement rates, at 55.40% and 55.00%, respectively. The positive argumentative bias of E2 was therefore generated by higher scores in three dimensions rather than by a general increase applied to the complete rubric.

Under Rubric B, E4 assigned higher mean scores than E3 for all four technical criteria. The largest difference occurred for Completeness, where the mean score increased from 1.336 to 1.920. This criterion also produced the largest technical MAE, at 0.704 points, and the lowest exact agreement, at 37.80%. By contrast, Clarity showed the closest agreement, with an MAE of 0.272 points and exact agreement of 78.20%. Accuracy and Terminology also showed comparatively high exact agreement rates, at 59.20% and 69.40%. The overall technical difference between E3 and E4 was therefore driven primarily by Completeness.

For argumentative responses under Rubric B, E4 assigned higher mean scores for Clarity, Coherence, and Originality, but a lower mean score for Dialecticality. Originality produced the largest positive difference between the two evaluators, increasing from 0.558 to 1.144, and also showed the largest MAE in this subset, at 0.586 points. Coherence produced the highest exact agreement rate, at 63.00%, followed by Clarity at 61.20%. Dialecticality followed the opposite direction, with a mean score of 1.256 for E3 and 0.992 for E4. Thus, the positive argumentative total-score bias of E4 did not correspond to uniformly higher scores across all rubric dimensions.

Taken together, the human evaluations exhibited both systematic differences in overall scoring level and criterion-specific differences in point allocation. The two evaluator pairs preserved a meaningful correspondence across responses, but their scores were not interchangeable at either the final-score or criterion level. These findings establish the human-agreement baseline for interpreting the rubric-matched model–human results.

3.2. Rubric-Matched Model–Human Alignment

The model-generated scores were compared with the human evaluations obtained under the corresponding rubric condition. This rubric-matched comparison is necessary because Rubrics A and B assign different point maxima to two criteria, and model scores generated under one allocation should not be evaluated against human scores obtained under the alternative allocation. Accordingly, model outputs generated under Rubric A were compared separately with E1 and E2, whereas the outputs generated under Rubric B were compared separately with E3 and E4.

The human evaluators were retained as separate reference points rather than being averaged into a single score. This preserves the evaluator-level variability observed within each rubric condition and avoids treating two non-identical human judgments as an interchangeable ground truth. Signed bias was calculated as the model score minus the human score; positive values therefore indicate that the model assigned higher scores, whereas negative values indicate lower model scores. MAE quantifies the magnitude of the response-level differences, while QWK measures agreement across the ordered 0–10 scale. Exact agreement represents the proportion of responses for which the model and the corresponding evaluator assigned the same final score.

All comparisons included 1000 matched responses, except for phi4-mini:3.8b under Rubric B, for which one structurally invalid output reduced the available sample to 999. In Table 6, H1/H2 denotes E1/E2 under Rubric A and E3/E4 under Rubric B.

Table 6. Overall rubric-matched model–human alignment.

Rubric	Model	Bias H1/H2	MAE H1/H2	QWK H1/H2	Exact H1/H2 (%)
A	qwen3:4b	−0.016/−0.685	1.364/1.901	0.704/0.634	22.30/18.40
A	qwen3:14b	0.730/0.061	1.442/1.699	0.706/0.708	22.00/19.10
A	gemma3:4b	1.908/1.239	2.174/2.179	0.501/0.540	12.90/14.60
A	gemma3:27b	1.792/1.123	1.998/1.977	0.547/0.605	16.50/17.00
A	llama3.1:8b	1.312/0.643	1.846/1.963	0.562/0.597	17.70/16.90
A	llama3.1:70b	0.169/−0.500	1.315/1.684	0.749/0.719	22.90/19.80
A	phi4-mini:3.8b	0.426/−0.243	1.654/2.313	0.451/0.414	18.60/10.50
A	phi4:14b	1.354/0.685	1.794/1.795	0.604/0.662	18.90/19.10
B	qwen3:4b	−0.691/−1.647	1.489/2.105	0.706/0.606	23.60/14.70
B	qwen3:14b	−0.126/−1.082	1.250/1.674	0.761/0.691	25.20/20.60
B	gemma3:4b	1.131/0.175	1.671/1.509	0.638/0.709	18.60/22.30
B	gemma3:27b	0.295/−0.661	1.235/1.505	0.750/0.736	28.10/19.60
B	llama3.1:8b	−0.420/−1.376	1.478/2.060	0.637/0.527	18.90/11.80
B	llama3.1:70b	−0.955/−1.911	1.425/2.209	0.697/0.555	23.70/12.50
B	phi4-mini:3.8b	−0.796/−1.752	1.925/2.480	0.349/0.299	11.51/11.91
B	phi4:14b	−0.140/−1.096	1.276/1.768	0.714/0.636	25.80/16.10

Under Rubric A, the direction and magnitude of the bias depended on the evaluator selected as reference. qwen3:4b was almost perfectly centered relative to E1, with a bias of −0.016 points, but underestimated E2 by 0.685 points. Conversely, qwen3:14b overestimated E1 by 0.730 points while differing from E2 by only 0.061 points.

Llama3.1:70b produced the strongest overall response-level alignment under Rubric A. It obtained the lowest MAE relative to both E1 and E2, at 1.315 and 1.684 points, respectively, together with the highest corresponding QWK values, at 0.749 and 0.719. qwen3:4b and qwen3:14b also showed comparatively strong alignment, although their relative performance differed between the two human references and across the individual metrics.

The two Gemma models assigned substantially higher scores than both Rubric A evaluators. Their biases ranged from 1.123 to 1.908 points, while their MAE values were close to or above two points. Their QWK values were also lower than those obtained by the Qwen models and llama3.1:70b, indicating that the disagreement was not limited to a uniform increase in mean scoring level. phi4-mini:3.8b showed a different pattern. Its QWK values were only 0.451 and 0.414. Thus, a moderate bias did not necessarily correspond to close response-level agreement.

Most models underestimated E4, with negative biases exceeding one point for qwen3:4b, qwen3:14b, llama3.1:8b, llama3.1:70b, phi4-mini:3.8b, and phi4:14b. gemma3:4b was the only model with a positive bias relative to E4, at 0.175 points. However, this small bias did not correspond to the lowest response-level error.

Gemma3:27b produced the lowest overall MAE under Rubric B relative to both E3 and E4, at 1.235 and 1.505 points, respectively. It also obtained QWK values of 0.750 and 0.736. qwen3:14b achieved the highest overall QWK relative to E3, at 0.761, while maintaining a similarly low MAE of 1.250 points. Relative to E4, gemma3:27b produced the highest QWK, whereas gemma3:4b yielded a similar MAE but a lower QWK. The models therefore differed according to whether alignment was characterized by average error, ordinal agreement, or signed bias.

Phi4-mini:3.8b showed the weakest overall alignment under Rubric B. Its MAE reached 1.925 points relative to E3 and 2.480 points relative to E4, while QWK decreased to 0.349 and 0.299. The model also produced the lowest exact-agreement rate relative to E3 and one of the lowest rates relative to E4. Its sample contained 999 responses because the single structurally invalid output identified during validation occurred for this model under Rubric B; the missing score was not imputed.

Exact final-score agreement remained limited across all models. Under Rubric A, the highest rates were 22.90% relative to E1 and 19.80% relative to E2, both obtained by llama3.1:70b. Under Rubric B, the maximum exact-agreement rate relative to E3 was 28.10% for gemma3:27b, whereas gemma3:4b reached the highest value relative to E4, at 22.30%. These values reinforce the distinction between exact score reproduction and broader ordinal agreement, as several models achieved comparatively high QWK values while matching the human score exactly in fewer than one-third of the responses.

The strongest model–human results were numerically comparable to the aggregate human–human agreement observed on the same 0–10 scale. Under Rubric A, the human–human MAE, QWK, and exact agreement were 1.827, 0.678, and 16.7%, respectively, whereas the best model–human values across E1 and E2 ranged from 1.315 to 1.684 for MAE, from 0.719 to 0.749 for QWK, and from 19.8% to 22.9% for exact agreement. Under Rubric B, the corresponding human–human values were 1.304, 0.783, and 28.3%, while the best model–human values across E3 and E4 ranged from 1.235 to 1.505, from 0.736 to 0.761, and from 22.3% to 28.1%, respectively. Thus, the strongest model–human correspondence was comparable to or greater than human–human agreement under Rubric A, whereas under Rubric B it was comparable for some evaluator–metric combinations but weaker for others.

The overall values may conceal differences between technical and argumentative responses. Figure 2 therefore presents MAE and QWK separately for the two response types and for each rubric-matched evaluator.

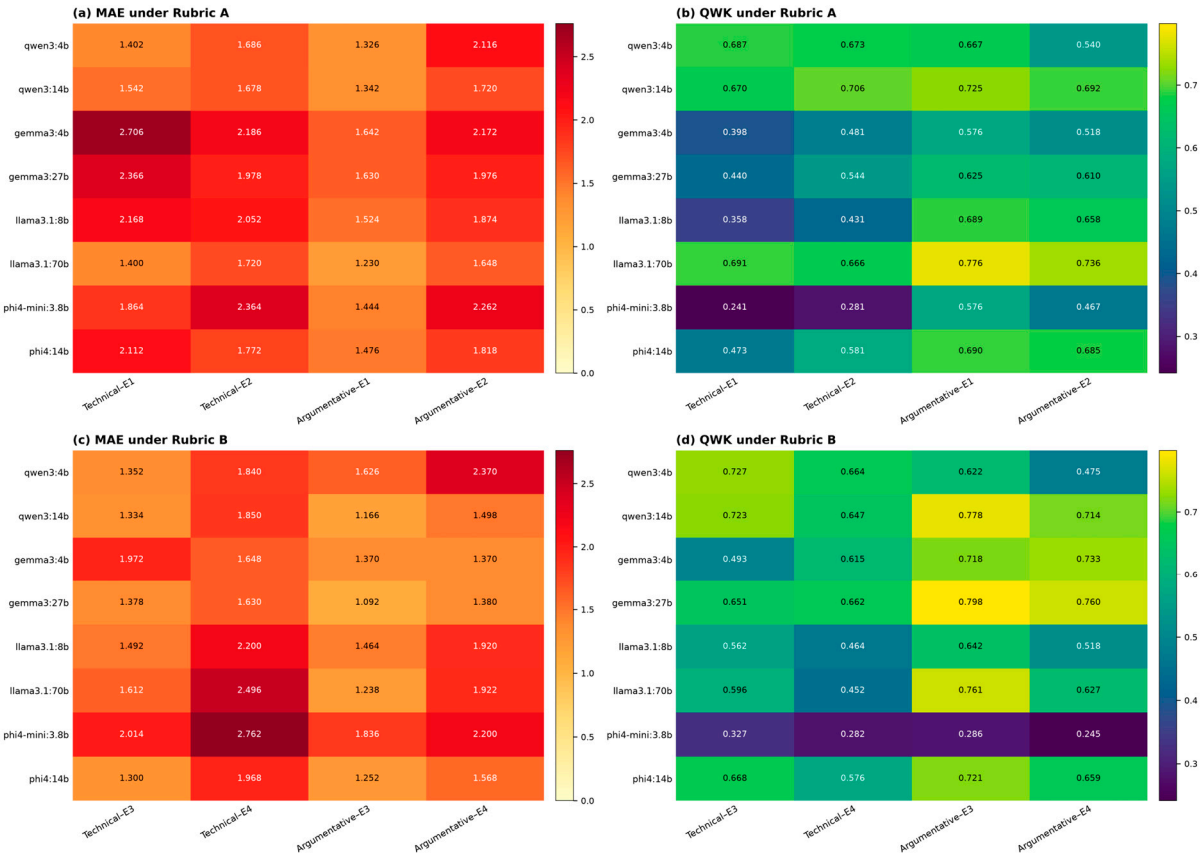


Figure 2. Rubric-matched model–human alignment by response type.

Under Rubric A, llama3.1:70b showed the most consistent alignment across response types. It obtained the lowest MAE and the highest QWK relative to E1 for both technical and argumentative responses. Relative to E2, qwen3:14b performed slightly better for technical

responses, with an MAE of 1.678 and QWK of 0.706, whereas llama3.1:70b showed the strongest argumentative alignment, with an MAE of 1.648 and QWK of 0.736. The model most closely aligned with one evaluator or response type was therefore not necessarily the closest under all matched comparisons.

For several models, argumentative responses produced stronger alignment than technical responses under Rubric A. This difference was particularly visible for llama3.1:8b relative to E1, for which QWK increased from 0.358 for technical responses to 0.689 for argumentative responses, and for gemma3:27b, for which the corresponding values increased from 0.440 to 0.625. The pattern was not universal, however. qwen3:4b showed a lower argumentative MAE relative to E1 but a substantially higher argumentative MAE relative to E2, demonstrating that the response-type difference also depended on the human evaluator used as reference.

Under Rubric B, gemma3:27b showed the strongest ordinal agreement for argumentative responses, with QWK values of 0.798 relative to E3 and 0.760 relative to E4. It also produced the lowest MAE relative to E3, at 1.092 points. Relative to E4, however, the lowest argumentative MAE was obtained by gemma3:4b, at 1.370 points, closely followed by gemma3:27b at 1.380 points. qwen3:14b also showed strong argumentative alignment, particularly relative to E3, where QWK reached 0.778.

For technical responses under Rubric B, the strongest model depended on the metric and evaluator. phi4:14b produced the lowest MAE relative to E3, whereas qwen3:4b obtained the highest technical QWK relative to both E3 and E4. The difference between these results again shows that the smallest absolute error and the strongest preservation of score ordering do not necessarily identify the same model.

The response-type analysis also identified models for which alignment changed substantially between assessment categories. qwen3:4b aligned more closely with the Rubric B evaluators on technical responses than on argumentative responses, particularly relative to E4, where QWK decreased from 0.664 to 0.475. In contrast, qwen3:14b, gemma3:27b, and llama3.1:70b generally produced stronger ordinal agreement for argumentative responses. phi4-mini:3.8b remained weak across both categories and reached its lowest QWK values for argumentative responses under Rubric B, at 0.286 relative to E3 and 0.245 relative to E4.

Taken together, the model–human comparisons show that alignment cannot be represented by a single evaluator-independent ranking. The observed results depended jointly on the active rubric, the individual human evaluator, the response type, and the metric used to characterize agreement. Some models reproduced a human mean closely while showing weaker response-level correspondence, whereas others maintained low MAE and high QWK despite a systematic directional bias. These results provide the rubric-matched reference required for examining how the same models modified their scores when the criterion weights were redistributed.

3.3. Total-Score Shifts Under Alternative Rubrics

Same-rubric repeatability varied substantially across models. Pooled across all models, A1–A2 and B1–B2 showed exact final-score agreement of 59.63% and 59.29%, MAE of 0.536 and 0.569 points, and QWK of 0.920 and 0.907, respectively. Signed bias (repeat minus original) was +0.028 points under Rubric A and +0.397 points under Rubric B. Across both rubrics, the two Qwen models achieved 89.30–95.80% exact final-score agreement, compared with 23.92–75.90% for the other models. Thus, identical prompts and generation parameters did not ensure identical final scores.

To determine whether the redistribution of criterion maxima affected the final scores assigned by the models, the Rubric A and Rubric B evaluations were paired for the same model and response. For each paired model–response evaluation, the total-score shift,

ΔT_{im}, was calculated by subtracting the Rubric A total score from the corresponding Rubric B total score, as specified in Equation (2).

Negative values therefore indicate a lower score under Rubric B, zero indicates an unchanged final score, and positive values indicate a higher score under Rubric B. Because both rubrics retained the same maximum total of 10 points, the observed shifts cannot be explained by a mechanical reduction in the number of available points. They may, however, include both rubric-related changes and run-to-run variability.

The analysis included 7999 complete and valid Rubric A–Rubric B pairs. Each model contributed 1000 paired responses, except phi4-mini:3.8b, for which the structurally invalid Rubric B output identified during validation reduced the available sample to 999. No score was imputed for the excluded response. Mean shifts and their 95% confidence intervals were estimated using 10,000 student-cluster bootstrap replicates. Table 7 reports the overall shift, its median and interquartile range, the mean shifts for technical and argumentative responses, and the proportions of responses for which the total score decreased, remained unchanged, or increased.

Table 7. Total-score shifts between Rubric A and Rubric B.

Model	Overall Mean ΔT [95% CI]	Median (IQR)	Technical Mean ΔT [95% CI]	Argumentative Mean ΔT [95% CI]	Decrease/Unchanged/Increase (%)
qwen3:4b	−0.182 [−0.258, −0.108]	0.000 (1.000)	−0.136 [−0.252, −0.026]	−0.228 [−0.322, −0.132]	31.10/46.90/22.00
qwen3:14b	−0.363 [−0.415, −0.312]	0.000 (1.000)	−0.278 [−0.346, −0.212]	−0.448 [−0.534, −0.364]	34.10/55.70/10.20
gemma3:4b	−0.284 [−0.350, −0.216]	0.000 (1.000)	−0.264 [−0.338, −0.188]	−0.304 [−0.412, −0.198]	34.20/52.10/13.70
gemma3:27b	−1.004 [−1.070, −0.938]	−1.000 (1.000)	−0.998 [−1.070, −0.924]	−1.010 [−1.102, −0.920]	73.70/24.70/1.60
llama3.1:8b	−1.239 [−1.346, −1.131]	−1.000 (2.000)	−1.594 [−1.690, −1.492]	−0.884 [−1.070, −0.694]	73.70/15.00/11.30
llama3.1:70b	−0.631 [−0.709, −0.557]	−1.000 (1.000)	−0.746 [−0.830, −0.666]	−0.516 [−0.620, −0.416]	51.90/41.10/7.00
phi4-mini:3.8b	−0.731 [−0.816, −0.646]	−1.000 (2.000)	−0.818 [−0.916, −0.724]	−0.644 [−0.812, −0.482]	55.26/28.43/16.32
phi4:14b	−1.001 [−1.089, −0.913]	−1.000 (2.000)	−1.362 [−1.440, −1.280]	−0.640 [−0.776, −0.508]	66.80/26.90/6.30

All eight models assigned lower mean total scores under Rubric B than under Rubric A, and the 95% bootstrap confidence interval for the mean shift remained entirely below zero for every model. The magnitude of the reduction nevertheless varied substantially. qwen3:4b showed the smallest overall change, with a mean shift of −0.182 points, followed by gemma3:4b at −0.284 and qwen3:14b at −0.363. At the other end of the distribution, llama3.1:8b produced the largest reduction, at −1.239 points, while gemma3:27b and phi4:14b each reduced their scores by approximately one point on average.

The median shift was zero for qwen3:4b, qwen3:14b, and gemma3:4b, whereas the remaining five models had a median shift of −1 point. The negative mean shifts therefore arose from different response-level distributions rather than from a uniform one-point reduction across all models and responses.

The directional proportions further distinguish the observed behaviors. Qwen3:14b and gemma3:4b retained the same total score in more than half of the responses, while qwen3:4b also showed a comparatively high unchanged proportion. Despite this stability, decreases remained more frequent than increases for all three models, producing negative average shifts.

In contrast, gemma3:27b and llama3.1:8b reduced the total score in nearly three-quarters of the paired evaluations, and phi4:14b also showed a predominantly negative response. These models therefore reacted to the alternative criterion allocation primarily by reducing the final score rather than by preserving or increasing it.

The overall results combine technical and argumentative responses, although the two assessment types redistribute points between different pairs of criteria. Figure 3 separates

the direction of the total-score shifts by response type to show whether the aggregated patterns were consistent across the two categories.

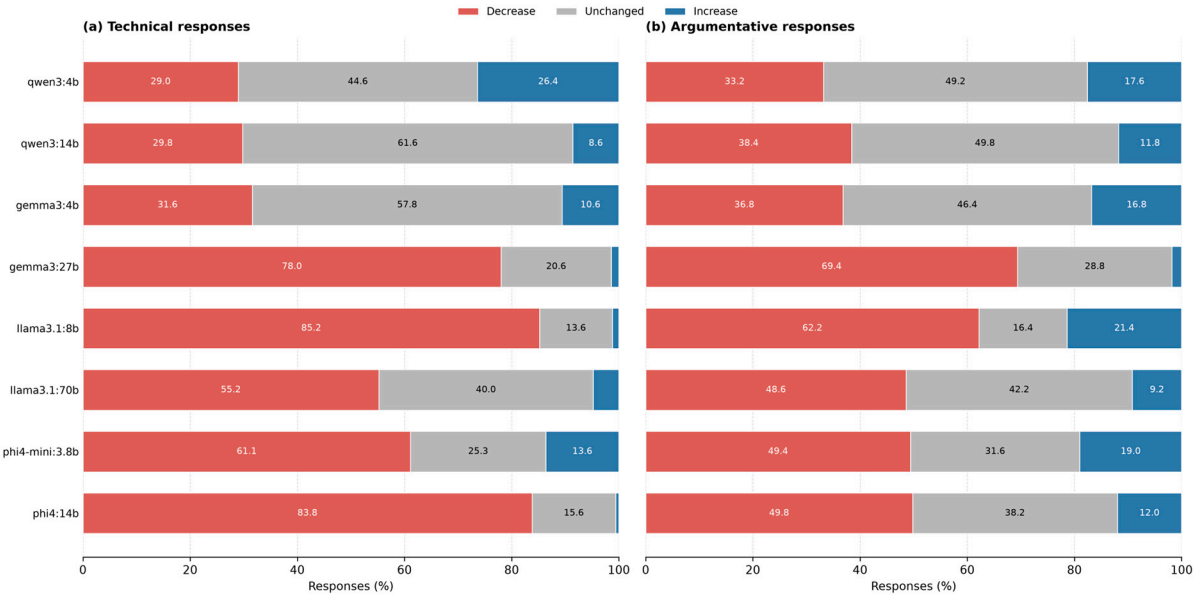


Figure 3. Direction of total-score shifts by model and response type.

For qwen3:4b, qwen3:14b, and gemma3:4b, unchanged scores represented the largest category for both response types. qwen3:14b preserved the technical score in 61.60% of the paired evaluations and the argumentative score in 49.80%, while gemma3:4b retained the corresponding scores in 57.80% and 46.40% of cases. qwen3:4b exhibited the most balanced distribution for technical responses: 29.00% decreased, 44.60% remained unchanged, and 26.40% increased. Its technical mean shift was consequently limited to −0.136 points. For argumentative responses, the decrease proportion rose to 33.20% and the increase proportion declined to 17.60%, producing a more negative mean shift of −0.228 points.

qwen3:14b and gemma3:4b also showed somewhat larger reductions for argumentative responses than for technical responses. For qwen3:14b, the mean shift changed from −0.278 for technical responses to −0.448 for argumentative responses, while the proportion of decreases increased from 29.80% to 38.40%. The corresponding mean shifts for gemma3:4b were −0.264 and −0.304. Although these differences were smaller than those observed for several other models, they show that the effect of the alternative rubric was not necessarily identical across response categories even among models that frequently preserved the original total score.

gemma3:27b showed the most consistent negative response across the two assessment types. Its mean shifts were −0.998 for technical responses and −1.010 for argumentative responses, with closely overlapping bootstrap intervals. The model reduced the technical score in 78.00% of cases and the argumentative score in 69.40%, while increases occurred in only 1.40% and 1.80%, respectively. The similarity of the mean changes indicates that its approximately one-point overall reduction was not driven predominantly by either response category.

The strongest response-type differences were observed for llama3.1:8b and phi4:14b. llama3.1:8b reduced technical scores by an average of 1.594 points, compared with 0.884 points for argumentative scores. Technical scores decreased in 85.20% of responses and increased in only 1.20%, whereas argumentative scores decreased in 62.20% and increased in 21.40%. Thus, the negative effect remained present for both assessment types but was considerably stronger and more uniform for technical responses.

A similar pattern occurred for phi4:14b. Its mean technical shift was −1.362 points, compared with −0.640 points for argumentative responses. The technical score decreased in 83.80% of cases and increased in only 0.60%. For argumentative responses, decreases accounted for 49.80%, unchanged scores for 38.20%, and increases for 12.00%. The aggregated reduction of approximately one point for this model therefore concealed a pronounced concentration of the effect in the technical subset.

llama3.1:70b and phi4-mini:3.8b also showed larger reductions for technical than for argumentative responses, although the differences were less pronounced. For llama3.1:70b, the mean shifts were −0.746 and −0.516 points, respectively. For phi4-mini:3.8b, they were −0.818 and −0.644 points. In both cases, technical responses contained higher proportions of decreases and lower proportions of increases than argumentative responses.

Taken together, the results show that all evaluated models assigned lower mean total scores under Rubric B than under Rubric A, but the magnitude and response-level structure of that change were model dependent. Some models preserved the same final score for approximately half or more of the responses, while others reduced scores in more than two-thirds of the paired evaluations. The differences between technical and argumentative responses further indicate that a model's aggregate shift may conceal assessment-type-specific behavior.

An unchanged total score does not, however, necessarily imply an unchanged criterion profile. A decrease in one criterion may be offset by an increase in another, leaving the final score unchanged despite a substantive redistribution of points.

3.4. Adaptation of the Affected Criteria

Total-score shifts do not indicate which criterion scores changed when the alternative point allocation was applied. The analysis was therefore extended to the two criteria whose maxima differed between Rubrics A and B. For technical responses, the maximum assigned to Clarity decreased from 3 to 2 points, whereas the maximum assigned to Completeness increased from 2 to 3 points. For argumentative responses, the corresponding redistribution involved Clarity and Dialecticality.

Criterion changes were examined on both the raw and normalized scales. Raw shifts represent the difference between the criterion scores assigned under Rubrics B and A and therefore show how the criterion's direct contribution to the 10-point final score changed. Normalized shifts compare the proportion of the rubric-specific maximum awarded to the response. The two measures are not interchangeable: a criterion may receive fewer raw points under a reduced maximum while receiving a larger proportion of the points available under that rubric.

Table 8 reports the mean shifts for the two affected criteria and the normalized affected-profile distance, D^{aff}. In the paired columns, C denotes Clarity and G denotes the criterion whose maximum increased: Completeness for technical responses and Dialecticality for argumentative responses. FPI denotes full-profile insensitivity, defined as an identical four-criterion score profile across the two rubrics, whereas CS denotes a compensating shift, in which the total score remained unchanged despite at least one criterion-level modification. Each model–response-type comparison contained 500 complete pairs, except for phi4-mini:3.8b on technical responses, for which 499 pairs were available.

Table 8. Adaptation of the criteria affected by the alternative point allocations.

Model and Response Type	Raw Δ C/G [95% CI]	Normalized Δ C/G [95% CI]	D^{aff} [95% CI]	FPI/CS (%)
qwen3:4b—Technical	−0.556 [−0.608, −0.504]/ 0.506 [0.436, 0.574]	0.102 [0.079, 0.124]/ 0.011 [−0.011, 0.032]	0.171 [0.162, 0.179]	18.20/26.40
qwen3:4b—Argumentative	−0.282 [−0.328, −0.238]/ 0.134 [0.096, 0.176]	0.155 [0.132, 0.178]/ 0.015 [0.003, 0.027]	0.147 [0.136, 0.158]	46.20/3.00
qwen3:14b—Technical	−0.546 [−0.604, −0.488]/ 0.386 [0.326, 0.448]	0.108 [0.089, 0.128]/ −0.046 [−0.062, −0.030]	0.150 [0.138, 0.161]	35.20/26.40
qwen3:14b—Argumentative	−0.442 [−0.500, −0.384]/ 0.216 [0.172, 0.260]	0.113 [0.095, 0.131]/ −0.018 [−0.035, −0.003]	0.154 [0.145, 0.164]	39.60/10.20
gemma3:4b—Technical	−0.824 [−0.860, −0.788]/ 0.882 [0.840, 0.924]	0.031 [0.017, 0.045]/ 0.008 [−0.006, 0.023]	0.085 [0.072, 0.099]	4.20/53.60
gemma3:4b—Argumentative	−0.492 [−0.548, −0.436]/ 0.192 [0.134, 0.252]	0.148 [0.129, 0.167]/ −0.010 [−0.030, 0.009]	0.169 [0.156, 0.182]	26.80/19.60
gemma3:27b—Technical	−0.954 [−0.978, −0.926]/ 0.342 [0.286, 0.400]	0.010 [0.004, 0.016]/ −0.058 [−0.073, −0.042]	0.082 [0.076, 0.088]	4.20/16.40
gemma3:27b—Argumentative	−0.938 [−0.970, −0.906]/ 0.148 [0.114, 0.184]	−0.018 [−0.030, −0.008]/ −0.042 [−0.058, −0.026]	0.083 [0.074, 0.092]	9.20/19.60
llama3.1:8b—Technical	−1.084 [−1.128, −1.042]/ 0.612 [0.560, 0.664]	−0.052 [−0.074, −0.031]/ 0.020 [0.004, 0.036]	0.128 [0.116, 0.140]	1.80/11.80
llama3.1:8b—Argumentative	−0.976 [−1.060, −0.890]/ 0.018 [−0.026, 0.064]	−0.090 [−0.116, −0.063]/ −0.032 [−0.049, −0.016]	0.127 [0.117, 0.137]	9.00/7.40
llama3.1:70b—Technical	−0.762 [−0.820, −0.704]/ 0.250 [0.210, 0.292]	0.014 [−0.008, 0.037]/ −0.059 [−0.075, −0.044]	0.174 [0.163, 0.184]	27.40/12.60
llama3.1:70b—Argumentative	−0.692 [−0.762, −0.618]/ 0.102 [0.064, 0.144]	−0.010 [−0.028, 0.008]/ −0.007 [−0.021, 0.007]	0.091 [0.085, 0.098]	21.60/20.60
phi4-mini:3.8b—Technical	−0.982 [−1.082, −0.880]/ 0.994 [0.922, 1.066]	−0.117 [−0.154, −0.079]/ 0.109 [0.086, 0.132]	0.275 [0.265, 0.286]	1.20/24.05
phi4-mini:3.8b—Argumentative	−1.064 [−1.186, −0.944]/ 0.014 [−0.062, 0.092]	−0.188 [−0.229, −0.148]/ −0.044 [−0.073, −0.014]	0.236 [0.221, 0.251]	24.20/7.40
phi4:14b—Technical	−0.952 [−0.984, −0.916]/ 0.132 [0.092, 0.176]	−0.008 [−0.019, 0.002]/ −0.111 [−0.125, −0.096]	0.102 [0.096, 0.109]	7.60/8.00
phi4:14b—Argumentative	−0.860 [−0.946, −0.774]/ 0.222 [0.176, 0.268]	−0.066 [−0.090, −0.041]/ −0.027 [−0.049, −0.004]	0.173 [0.163, 0.183]	20.20/18.00

On the raw scale, the Clarity score decreased for every model and response type, and all corresponding confidence intervals remained below zero. The magnitude of the decrease varied considerably. For technical responses, the mean reduction ranged from 0.546 points for qwen3:14b to 1.084 points for llama3.1:8b. For argumentative responses, the smallest decrease was observed for qwen3:4b, at 0.282 points, whereas phi4-mini:3.8b reduced Clarity by 1.064 points on average.

The criterion whose maximum increased generally received more raw points under Rubric B. Completeness increased significantly for every model on technical responses, with mean changes ranging from 0.132 points for phi4:14b to 0.994 points for phi4-mini:3.8b. Dialecticality also increased for most models, although the changes were considerably smaller. For llama3.1:8b and phi4-mini:3.8b, the mean increases were only 0.018 and 0.014 points, respectively, and their confidence intervals included zero.

The balance between the two raw changes differed across models. gemma3:4b showed a nearly symmetric technical redistribution, with Clarity decreasing by 0.824 points and Completeness increasing by 0.882 points. phi4-mini:3.8b showed an even closer balance, at −0.982 and 0.994 points. These patterns are consistent with the high compensating-shift rates observed for the two models, because changes in opposite directions frequently cancelled at the total-score level. In contrast, for gemma3:27b, llama3.1:8b, and phi4:14b, the

reduction in technical Clarity substantially exceeded the gain in Completeness, contributing to their more negative total-score shifts.

The normalized results provide a different view of the same changes. qwen3:4b and qwen3:14b awarded fewer raw Clarity points under Rubric B but increased the proportion of the reduced Clarity maximum awarded to technical responses by 0.102 and 0.108, respectively. A similar pattern appeared for argumentative Clarity in qwen3:4b, qwen3:14b, and gemma3:4b. These models therefore reduced the criterion's numerical contribution to the final score without evaluating the responses more strictly relative to the new two-point scale.

Other models reduced both the raw Clarity score and its normalized proportion. This pattern was particularly pronounced for phi4-mini:3.8b, whose normalized Clarity shifts were -0.117 for technical responses and -0.188 for argumentative responses. llama3.1:8b and phi4:14b also showed negative normalized changes for both response types. For these models, the Clarity reduction was not merely the arithmetic consequence of the lower maximum; it was accompanied by a lower share of the available criterion points.

The converse occurred for the criteria whose maxima increased. A positive raw shift did not necessarily imply an increased normalized score. qwen3:14b, gemma3:27b, llama3.1:70b, and phi4:14b awarded more raw Completeness points under Rubric B but a smaller proportion of the new three-point maximum. The same pattern was visible for argumentative Dialecticality in most models. qwen3:4b was the only model for which the normalized Dialecticality shift was clearly positive, at 0.015, whereas several models showed small but consistently negative normalized shifts despite assigning slightly more raw points.

The affected-profile distance, D^{aff}, summarizes the magnitude of the normalized changes without indicating whether they occurred in the expected direction. The highest values were obtained by phi4-mini:3.8b for both technical and argumentative responses, at 0.275 and 0.236, respectively. By contrast, gemma3:27b produced the smallest distances, at 0.082 and 0.083. These values should therefore be interpreted as the extent of criterion-profile modification rather than as a direct measure of successful rubric adaptation. A large distance may reflect a strong concordant redistribution, but it may also include changes in the opposite direction.

The FPI and CS results further distinguish unchanged scoring from internal redistribution. The most prominent case was gemma3:4b on technical responses. Only 4.20% of its complete criterion profiles remained unchanged, whereas 53.60% of the responses showed compensating shifts. Thus, more than half of the technical responses retained the same final score while undergoing at least one criterion-level modification. Compensating shifts were also frequent for qwen3:4b and qwen3:14b on technical responses, reaching 26.40% for both models, and for phi4-mini:3.8b, at 24.05%.

The pattern was different for argumentative responses. qwen3:4b showed the highest full-profile insensitivity rate, at 46.20%, together with the lowest compensating-shift rate, at 3.00%. qwen3:14b also retained the entire criterion profile in 39.60% of cases. These findings indicate that unchanged argumentative totals for the Qwen models more frequently reflected genuinely unchanged scoring profiles, whereas unchanged technical totals often concealed offsetting criterion-level changes.

The direction of the raw shifts was examined more directly by classifying each paired evaluation as fully concordant, partially concordant, unchanged on both affected criteria, or discordant. Full concordance required both a reduction in Clarity and an increase in the criterion whose maximum increased. Partial concordance required one expected change while the other affected criterion remained unchanged. A case was classified as

discordant when Clarity increased or the gaining criterion decreased. Figure 4 presents the resulting distributions.

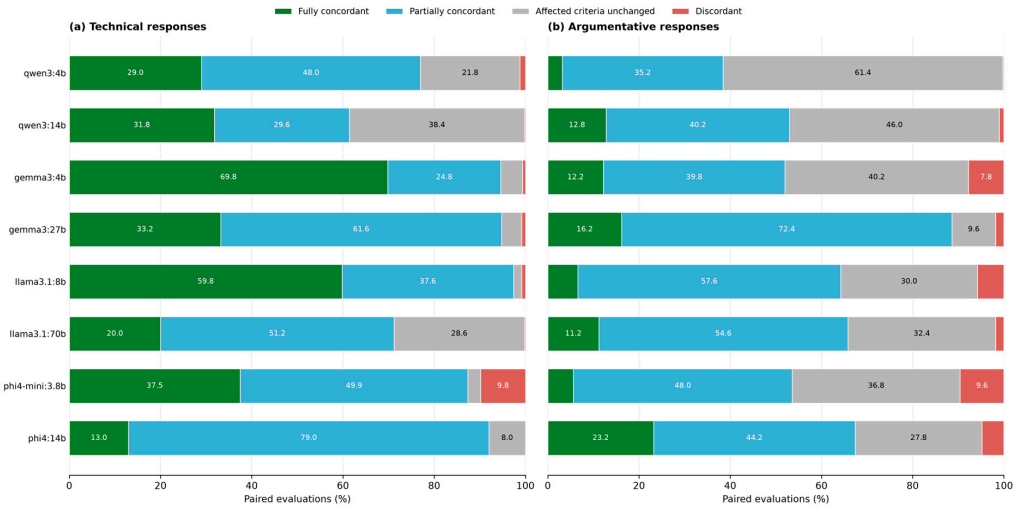


Figure 4. Directional adaptation of the affected criteria by model and response type.

For technical responses, the combined proportion of fully and partially concordant cases exceeded 60% for every model. It ranged from 61.40% for qwen3:14b to 97.40% for llama3.1:8b. gemma3:4b exhibited the highest fully concordant rate, at 69.80%, followed by llama3.1:8b at 59.80%. For these models, both affected criteria frequently moved in the direction implied by the revised allocation.

Several other models adapted predominantly through a change in only one affected criterion. Partial concordance accounted for 79.00% of the technical responses evaluated by phi4:14b, 61.60% for gemma3:27b, and 51.20% for llama3.1:70b. Their relatively low fully concordant rates therefore did not imply an absence of rubric sensitivity; rather, the redistribution was usually localized to one of the two affected dimensions.

The affected criteria remained jointly unchanged in 38.40% of the technical responses for qwen3:14b and 28.60% for llama3.1:70b. For gemma3:4b, gemma3:27b, llama3.1:8b, and phi4-mini:3.8b, this category represented fewer than 5% of cases. Discordant technical shifts were uncommon for seven of the eight models, ranging from 0% for phi4:14b to 1.20% for qwen3:4b. The exception was phi4-mini:3.8b, for which 9.82% of technical pairs contained at least one shift opposite to the expected direction.

Directional adaptation was more heterogeneous for argumentative responses. The combined concordant proportion ranged from 38.40% for qwen3:4b to 88.60% for gemma3:27b. gemma3:27b adapted primarily through partial concordance, which accounted for 72.40% of its argumentative responses. Partial concordance was also the largest category for llama3.1:8b, llama3.1:70b, phi4-mini:3.8b, and phi4:14b.

For qwen3:4b, the affected argumentative criteria remained unchanged in 61.40% of cases, while only 3.20% were fully concordant. qwen3:14b and gemma3:4b also showed comparatively high unchanged proportions, at 46.00% and 40.20%, respectively. These results explain why the argumentative adaptations of these models were weaker than their technical adaptations despite the consistent mean reduction in Clarity.

Discordant argumentative changes remained a minority but were more frequent than in the technical subset for several models. The highest rates occurred for phi4-mini:3.8b, at 9.60%, and gemma3:4b, at 7.80%. llama3.1:8b and phi4:14b showed discordant rates of 5.80% and 4.80%, respectively. Consequently, the large affected-profile distances observed for phi4-mini:3.8b cannot be interpreted solely as strong conformity to the new point allocation, because part of that movement occurred in the opposite direction.

Taken together, the criterion-level results show that the models generally recognized the redistribution of the affected criteria, but they did so through different mechanisms. Some models changed both criteria in the expected direction, whereas others modified only one dimension or retained both affected scores. The contrast between raw and normalized shifts also demonstrates that assigning fewer or more raw points does not necessarily imply a stricter or more generous evaluation relative to the available criterion maximum.

Most importantly, the analysis confirms that an unchanged final score cannot be treated as evidence of rubric insensitivity. In several model–response-type combinations, particularly gemma3:4b on technical responses, substantial criterion-level redistribution occurred without changing the total score. The localization of these adaptations was assessed by examining whether score changes extended to criteria whose definitions and point allocations were identical under both rubrics.

3.5. Stability of the Unaffected Criteria

The models frequently modified the criteria whose point maxima changed between Rubrics A and B. A localized adaptation would preserve the criteria whose definitions and maximum scores remained identical across the two conditions. The present analysis therefore examined Accuracy and Terminology for technical responses and Coherence and Originality for argumentative responses.

For each unaffected criterion, the raw shift was calculated as the Rubric B score minus the corresponding Rubric A score. Because the criterion maxima were unchanged, the raw and normalized shifts convey the same directional information, differing only by the constant criterion maximum. Exact criterion stability indicates the proportion of paired evaluations for which the criterion score remained identical. Joint stability requires both unaffected criteria to remain unchanged, whereas spillover denotes a change in at least one of them. The normalized unaffected-profile distance, D^{unaff}, summarizes the absolute magnitude of the changes across the two unaffected criteria.

In Table 9, U1/U2 denotes Accuracy/Terminology for technical responses and Coherence/Originality for argumentative responses. Each model–response-type comparison contained 500 complete pairs, except for phi4-mini:3.8b on technical responses, for which 499 pairs were available.

Table 9. Stability of the criteria unaffected by the alternative point allocations.

Model and Response Type	Raw Δ U1/U2 [95% CI]	Exact Stability U1/U2 (%)	D_unaff [95% CI]	Joint Stability/Spillover (%)
qwen3:4b—Technical	−0.050 [−0.082, −0.020]/ −0.036 [−0.064, −0.008]	90.80/90.80	0.039 [0.030, 0.047]	84.60/15.40
qwen3:4b—Argumentative	−0.140 [−0.176, −0.104]/ 0.060 [0.034, 0.088]	84.00/90.40	0.051 [0.043, 0.059]	76.20/23.80
qwen3:14b—Technical	−0.042 [−0.058, −0.026]/ −0.076 [−0.102, −0.050]	95.80/89.20	0.034 [0.027, 0.042]	85.20/14.80
qwen3:14b—Argumentative	−0.220 [−0.260, −0.182]/ −0.002 [−0.028, 0.026]	78.00/91.40	0.058 [0.048, 0.069]	72.40/27.60
gemma3:4b—Technical	−0.216 [−0.258, −0.174]/ −0.106 [−0.142, −0.070]	76.00/82.20	0.085 [0.074, 0.096]	63.00/37.00
gemma3:4b—Argumentative	0.060 [−0.024, 0.144]/ −0.064 [−0.092, −0.036]	50.80/88.60	0.114 [0.104, 0.125]	43.60/56.40
gemma3:27b—Technical	−0.104 [−0.130, −0.078]/ −0.282 [−0.326, −0.238]	88.80/70.60	0.092 [0.081, 0.104]	64.40/35.60
gemma3:27b—Argumentative	−0.104 [−0.152, −0.056]/ −0.116 [−0.150, −0.084]	76.40/86.40	0.073 [0.061, 0.086]	73.40/26.60

Table 9. *Cont.*

Model and Response Type	Raw Δ U1/U2 [95% CI]	Exact Stability U1/U2 (%)	D_unaff [95% CI]	Joint Stability/Spillover (%)
llama3.1:8b—Technical	−0.176 [−0.234, −0.122]/ −0.946 [−1.026, −0.866]	85.20/31.00	0.267 [0.248, 0.284]	24.80/75.20
llama3.1:8b—Argumentative	0.122 [0.020, 0.220]/ −0.048 [−0.092, −0.006]	43.80/86.40	0.134 [0.123, 0.146]	35.80/64.20
llama3.1:70b—Technical	−0.076 [−0.102, −0.052]/ −0.158 [−0.198, −0.118]	92.40/77.80	0.068 [0.057, 0.081]	74.00/26.00
llama3.1:70b—Argumentative	−0.024 [−0.080, 0.032]/ 0.098 [0.062, 0.136]	74.80/87.00	0.074 [0.064, 0.085]	64.40/35.60
phi4-mini:3.8b—Technical	−0.749 [−0.845, −0.654]/ −0.080 [−0.110, −0.052]	57.72/91.58	0.148 [0.130, 0.166]	51.30/48.70
phi4-mini:3.8b—Argumentative	0.286 [0.232, 0.336]/ 0.120 [0.084, 0.158]	58.80/85.00	0.110 [0.098, 0.122]	52.20/47.80
phi4:14b—Technical	−0.170 [−0.214, −0.128]/ −0.372 [−0.422, −0.324]	80.60/62.00	0.128 [0.116, 0.140]	47.40/52.60
phi4:14b—Argumentative	0.026 [−0.034, 0.086]/ −0.028 [−0.050, −0.006]	69.40/92.80	0.069 [0.060, 0.079]	64.80/35.20

For technical responses, all eight models produced negative mean shifts in both unaffected criteria, and every corresponding confidence interval remained below zero. The redistribution of points between Clarity and Completeness was therefore accompanied by lower mean scores not only in the affected dimensions but also in Accuracy and Terminology. The magnitude of this spillover nevertheless differed substantially across models.

The two Qwen models showed the highest technical stability. qwen3:14b preserved both unaffected criteria in 85.20% of the paired evaluations, closely followed by qwen3:4b at 84.60%. Their individual criterion scores were also highly stable: Accuracy remained unchanged in 95.80% and 90.80% of cases, respectively, while Terminology remained unchanged in 89.20% and 90.80%. Their D^{unaff} values, at 0.034 and 0.039, were the lowest among the technical comparisons. Although their mean shifts were consistently negative, the absolute changes were small and occurred in a limited subset of responses.

llama3.1:70b also showed comparatively strong technical localization. Both unaffected criteria remained unchanged in 74.00% of responses, and its unaffected-profile distance was 0.068. gemma3:4b and gemma3:27b exhibited intermediate levels of stability, with joint stability rates of 63.00% and 64.40%, respectively. For gemma3:27b, Terminology was less stable than Accuracy, remaining unchanged in 70.60% of cases compared with 88.80% for Accuracy.

The strongest technical spillover occurred for llama3.1:8b. Its joint stability rate was only 24.80%, meaning that at least one unaffected criterion changed in 75.20% of the paired evaluations. Accuracy remained unchanged in 85.20% of cases, but Terminology was stable in only 31.00%. The mean Terminology shift reached −0.946 points, with a confidence interval entirely below zero, while D^{unaff} reached 0.267, the largest unaffected-profile distance observed in the analysis.

The technical spillover produced by phi4-mini:3.8b was concentrated in the other unaffected dimension. Accuracy decreased by an average of 0.749 points and remained unchanged in only 57.72% of the paired evaluations, whereas Terminology was stable in 91.58%. Its joint stability rate was 51.30%, and D^{unaff} reached 0.148. phi4:14b also showed substantial technical spillover, but the larger component occurred in Terminology, whose mean score decreased by 0.372 points and remained unchanged in only 62.00% of cases.

The argumentative results were more heterogeneous in both direction and magnitude. Unlike the uniformly negative technical shifts, some models increased one or both

unaffected argumentative criteria. qwen3:4b reduced Coherence by 0.140 points while increasing Originality by 0.060 points. llama3.1:8b increased Coherence by 0.122 points while reducing Originality by 0.048 points. phi4-mini:3.8b increased both unaffected criteria, with mean changes of 0.286 for Coherence and 0.120 for Originality. These changes indicate that spillover was not limited to a general downward adjustment of the entire scoring profile.

qwen3:4b showed the highest argumentative joint stability, preserving both unaffected criteria in 76.20% of the responses. gemma3:27b and qwen3:14b followed at 73.40% and 72.40%, respectively. Their D^{unaff} values were also comparatively low, ranging from 0.051 for qwen3:4b to 0.073 for gemma3:27b. For these models, the majority of criterion changes remained localized to the dimensions whose maxima were altered.

The lowest argumentative joint stability was observed for llama3.1:8b, at 35.80%. Coherence remained unchanged in only 43.80% of cases, compared with 86.40% for Originality, resulting in a spillover rate of 64.20% and a D^{unaff} value of 0.134. gemma3:4b also showed substantial argumentative spillover. Both unaffected criteria remained unchanged in only 43.60% of responses, while Coherence alone was stable in 50.80%. However, the mean Coherence shift was only 0.060 points and its confidence interval included zero.

The gemma3:4b result illustrates an important distinction between average change and response-level stability. A mean shift close to zero does not imply that most individual scores remained unchanged. Positive and negative changes may offset one another in the aggregate while still indicating frequent response-level revisions. Exact stability and joint spillover are therefore necessary complements to the mean criterion shift.

phi4-mini:3.8b showed a similar contrast. Its argumentative Coherence and Originality means increased, but joint stability was only 52.20%, corresponding to spillover in 47.80% of the responses. llama3.1:70b and phi4:14b showed moderate argumentative spillover rates of 35.60% and 35.20%, respectively. In both cases, Coherence was less stable than Originality, although the mean Coherence confidence intervals included zero.

Figure 5 decomposes the joint spillover rate according to whether the change involved only the first unaffected criterion, only the second, or both simultaneously.

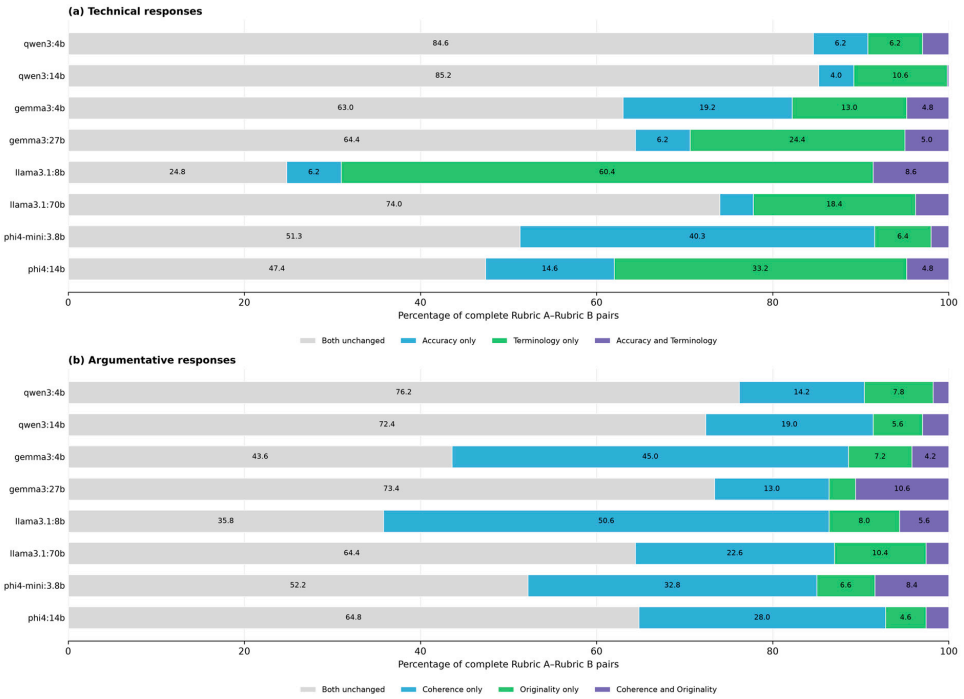


Figure 5. Localization of changes in the unaffected criteria by model and response type.

For the technical responses, the low joint stability of llama3.1:8b was driven predominantly by Terminology. Terminology alone changed in 60.40% of responses, whereas Accuracy alone changed in only 6.20%, and both criteria changed in 8.60%. The technical spillover of phi4:14b was similarly concentrated in Terminology, which changed alone in 33.20% of the responses. For gemma3:27b and llama3.1:70b, Terminology-only changes accounted for 24.40% and 18.40%, respectively.

By contrast, phi4-mini:3.8b modified Accuracy alone in 40.28% of its technical responses, while Terminology alone changed in only 6.41%. gemma3:4b showed a more balanced distribution: Accuracy alone changed in 19.20% of cases, Terminology alone in 13.00%, and both criteria in 4.80%. The two Qwen models remained comparatively stable, with all individual spillover categories below 11%.

For argumentative responses, the largest spillover component was generally Coherence. Coherence alone changed in 50.60% of the llama3.1:8b evaluations and 45.00% of the gemma3:4b evaluations. The corresponding proportions were 32.80% for phi4-mini:3.8b, 28.00% for phi4:14b, and 22.60% for llama3.1:70b. Originality-only changes were below 11% for every model.

gemma3:27b differed from this general pattern because simultaneous changes in Coherence and Originality accounted for 10.60% of the argumentative responses, the highest joint-change proportion in that subset. Nevertheless, 73.40% of its profiles remained unchanged on both unaffected criteria. For qwen3:4b and qwen3:14b, the combined rate of Originality-only and simultaneous changes remained below 10%, reinforcing their comparatively localized adaptation.

Comparison of the affected- and unaffected-profile distances showed that the changes were predominantly larger in the criteria whose maxima were modified. This localization was particularly clear for the Qwen models and phi4-mini:3.8b, for which D^{aff} substantially exceeded D^{unaff} in both response categories. The same relationship held for most of the remaining model–response-type combinations.

Several exceptions indicate broader reorganization of the score profile. For gemma3:27b on technical responses, D^{unaff} was 0.092 compared with $D^{aff} = 0.082$. For phi4:14b on technical responses, the corresponding values were 0.128 and 0.102. llama3.1:8b showed the clearest departure from localized adaptation: its technical D^{unaff} was 0.267, more than twice its D^{aff} of 0.128, while its argumentative D^{unaff} of 0.134 also slightly exceeded its D^{aff} of 0.127.

These cases show that cross-rubric score variation can also occur in dimensions that were formally unchanged. Such variation cannot be identified from total-score shifts alone and would be obscured by analyses restricted to the affected criteria.

Same-rubric comparisons showed that changes in the unaffected criteria also occurred without rubric redistribution. For the two Qwen models, joint stability across technical and argumentative responses was 93.60–97.60% within rubrics, compared with 72.40–85.20% between A1 and B1. In contrast, phi4-mini:3.8b on technical responses showed joint stability of 4.61% for A1–A2 and 3.21% for B1–B2, compared with 51.30% for A1–B1. Its corresponding normalized unaffected-profile distances were 0.367, 0.371, and 0.148, respectively. Across all models and response types, the A1–B1 normalized distance was numerically larger than both same-rubric baselines in 12 of the 16 comparisons. Cross-rubric changes in unaffected criteria therefore cannot be attributed exclusively to rubric redistribution and must be interpreted alongside model- and response-type-specific run-to-run variability.

3.6. Cross-Model and Response-Type Differences

Substantial variation was observed among models in the magnitude of the total-score shift, the modification of the affected criteria, and the spillover into unaffected dimensions. These differences were evaluated using student-level aggregates to preserve the dependence among the five responses contributed by each student within each response type. For every model, the mean total-score shift, affected-profile distance, and unaffected-profile distance were calculated separately across the technical and argumentative responses of each student.

Cross-model comparisons required complete student-level aggregates for all eight models. The technical analyses therefore included 99 students because the missing phi4-mini:3.8b evaluation prevented construction of a complete five-response technical aggregate for one student. The argumentative analyses included all 100 students. No missing score was imputed. Friedman tests were used to determine whether the eight models differed overall for each outcome and response type, and the resulting six p-values were adjusted jointly using the Holm procedure. The Friedman test statistics, Holm-adjusted p-values, and Kendall's W values are reported in Table 10.

Table 10. Overall cross-model differences in rubric adaptation.

Response Type	Outcome	Friedman $\chi^2(7)$	Adjusted p	Kendall's W
Technical	Mean total-score shift	401.937	<0.001	0.580
Technical	Affected-profile distance	390.359	<0.001	0.563
Technical	Unaffected-profile distance	349.317	<0.001	0.504
Argumentative	Mean total-score shift	103.058	<0.001	0.147
Argumentative	Affected-profile distance	325.422	<0.001	0.465
Argumentative	Unaffected-profile distance	161.612	<0.001	0.231

All six cross-model tests remained significant after Holm adjustment. For technical responses, the three outcomes produced Kendall's W values between 0.504 and 0.580, indicating pronounced differentiation among the models in both the direction of the total-score change and the extent to which the criterion profile was reorganized. The largest value was observed for the mean total-score shift, followed closely by the affected-profile distance. The unaffected-profile distance also varied strongly across models, confirming that spillover was not distributed uniformly across the evaluated systems.

The argumentative comparisons showed a different pattern. Cross-model differentiation remained substantial for the affected-profile distance, with $W = 0.465$, but was lower for the unaffected-profile distance, at $W = 0.231$. The smallest value occurred for the mean total-score shift, at $W = 0.147$. Thus, the models differed more consistently in how they redistributed criterion-level scores than in the aggregate reduction in argumentative totals.

Because significant cross-model differences were found for all six combinations of outcome and response type, paired Wilcoxon signed-rank comparisons were subsequently performed for the 28 model pairs within each combination. Holm adjustment was applied separately within each of the six post hoc families. Figure 6 presents the resulting comparisons. The lower-triangular matrices indicate the direction of the median student-level difference as row model minus column model. Negative values represent a lower total-score shift or smaller profile distance for the row model, whereas positive values represent a higher value. Nonsignificant comparisons are marked as *ns*.

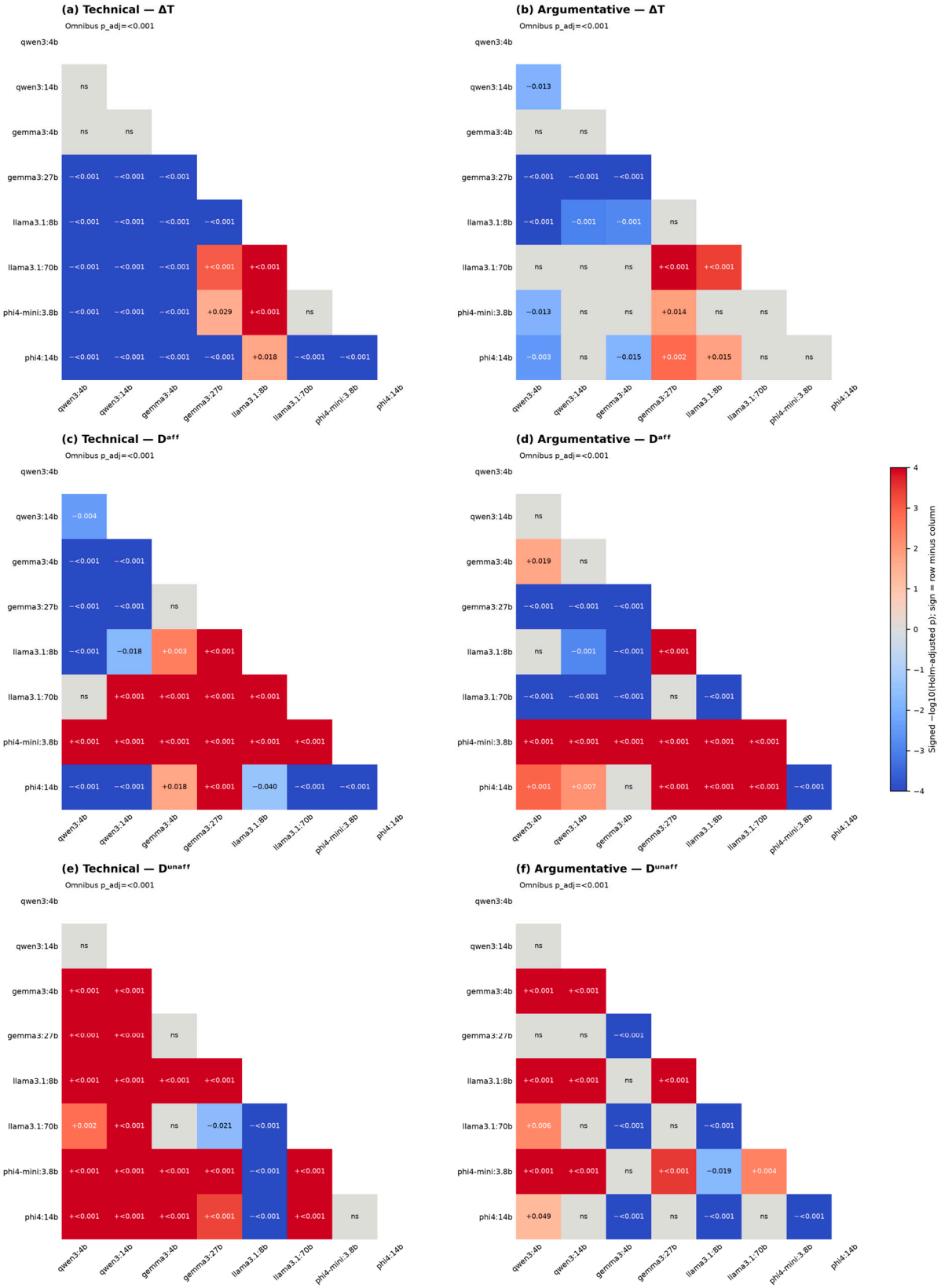


Figure 6. Holm-adjusted pairwise model differences in rubric adaptation.

The technical total-score shifts showed the clearest pairwise separation, with 24 of the 28 model comparisons remaining significant after Holm adjustment. The student-level median shifts were close for qwen3:4b, qwen3:14b, and gemma3:4b, at 0.0, −0.2, and −0.2 points, respectively, and none of the three pairwise contrasts among these models

was significant. By contrast, llama3.1:8b and phi4:14b showed median technical shifts of −1.6 and −1.4 points and differed significantly from the models with smaller reductions.

gemma3:27b produced a median technical shift of −1.0 points, while llama3.1:70b and phi4-mini:3.8b both had median shifts of −0.8 points. The difference between llama3.1:70b and phi4-mini:3.8b was not significant, although both differed from most models with smaller total-score reductions. These findings confirm that the negative mean shifts were not merely a common response shared by all systems; their magnitude differed consistently across students and model families.

Technical affected-profile distances differed in 26 of the 28 pairwise comparisons. phi4-mini:3.8b showed the largest student-level median distance, at 0.283, and differed significantly from every other model. qwen3:4b and llama3.1:70b had similar median distances, at 0.167 and 0.183, and their pairwise difference was not significant. gemma3:4b and gemma3:27b both had median values of 0.083 and were also statistically indistinguishable. The large number of significant comparisons demonstrates that models differed not only in whether their totals changed, but also in the extent to which the two affected criterion scores were reorganized.

The technical unaffected-profile distance also produced 24 significant pairwise comparisons. llama3.1:8b had the highest median value, at 0.250, and differed significantly from the other models, consistent with its extensive Terminology spillover. qwen3:4b and qwen3:14b had the lowest medians, both at 0.033, and did not differ significantly from one another. phi4-mini:3.8b and phi4:14b showed higher median distances, at 0.167 and 0.133, respectively, but their difference was not significant. The pairwise results therefore confirm the contrast between models that largely localized their adaptation and those that modified a broader portion of the technical rubric.

Pairwise separation was less extensive for the argumentative total-score shift, for which 15 of the 28 comparisons remained significant. gemma3:27b and llama3.1:8b both had median student-level shifts of −1.0 points and did not differ significantly from one another. Their reductions were generally larger than those of qwen3:4b, qwen3:14b, gemma3:4b, and llama3.1:70b, whose medians ranged from −0.2 to −0.4 points. phi4-mini:3.8b and phi4:14b occupied an intermediate position, both with median shifts of −0.7 points. The larger number of nonsignificant comparisons is consistent with the lower argumentative Kendall's W for this outcome.

Argumentative affected-profile distances remained strongly differentiated, with 23 of the 28 pairwise comparisons reaching significance. phi4-mini:3.8b produced the largest median distance, at 0.233, and differed significantly from all other models. gemma3:27b and llama3.1:70b had the smallest medians, both at 0.083, and did not differ significantly. qwen3:4b, qwen3:14b, gemma3:4b, llama3.1:8b, and phi4:14b occupied intermediate positions, although several adjacent comparisons within this group were not significant.

For the argumentative unaffected-profile distance, 18 of the 28 comparisons remained significant. llama3.1:8b had the largest median, at 0.133, followed by gemma3:4b and phi4-mini:3.8b, both at 0.100. qwen3:4b and qwen3:14b had the smallest medians, both at 0.050. Several models with median values between 0.067 and 0.100 were statistically indistinguishable, showing that the argumentative spillover patterns formed less clearly separated groups than the corresponding technical results.

The effect of response type was then evaluated within each model. Technical and argumentative student-level aggregates were compared using paired Wilcoxon signed-rank tests. The difference was defined as Technical minus Argumentative. A negative value for ΔT therefore indicates a larger score reduction for technical responses, whereas a positive value for either profile distance indicates greater technical criterion-level modification. Holm adjustment was applied jointly across the 24 within-model comparisons. All models

contributed 100 paired student aggregates, except phi4-mini:3.8b, for which 99 complete aggregates were available. The median technical–argumentative differences in total-score shift, affected-profile distance, and unaffected-profile distance, together with the Holm-adjusted p-values, are reported in Table 11.

Table 11. Within-model differences between technical and argumentative responses.

Model	ΔT, T − A Median Difference [p_adj]	D_aff, T − A Median Difference [p_adj]	D_unaff, T − A Median Difference [p_adj]
qwen3:4b	0.200 [1.000]	0.050 [0.024]	−0.017 [0.901]
qwen3:14b	0.200 [0.090]	−0.008 [1.000]	0.000 [0.007]
gemma3:4b	0.000 [1.000]	−0.083 [<0.001]	−0.025 [0.003]
gemma3:27b	0.000 [1.000]	0.000 [1.000]	0.017 [0.164]
llama3.1:8b	−0.600 [<0.001]	−0.017 [1.000]	0.133 [<0.001]
llama3.1:70b	−0.200 [0.002]	0.083 [<0.001]	0.000 [1.000]
phi4-mini:3.8b	−0.200 [0.882]	0.033 [0.003]	0.050 [0.010]
phi4:14b	−0.600 [<0.001]	−0.067 [<0.001]	0.050 [<0.001]

After adjustment, the technical–argumentative difference in the total-score shift was significant for three models. llama3.1:8b and phi4:14b each showed a median difference of −0.600 points, confirming substantially larger score reductions for technical responses. llama3.1:70b also showed a significantly more negative technical shift, although the median difference was smaller, at −0.200 points. The remaining five models did not show a significant response-type difference in their total-score change.

The affected-profile distance differed significantly between response types for five models. qwen3:4b, llama3.1:70b, and phi4-mini:3.8b showed larger technical distances, with median Technical-minus-Argumentative differences of 0.050, 0.083, and 0.033, respectively. In contrast, gemma3:4b and phi4:14b showed larger argumentative affected-profile distances, reflected by median differences of −0.083 and −0.067. No significant difference was observed for qwen3:14b, gemma3:27b, or llama3.1:8b.

The unaffected-profile distance also differed significantly for five models. llama3.1:8b showed the largest technical excess, with a median difference of 0.133, consistent with the strong Terminology spillover observed in its technical evaluations. phi4-mini:3.8b and phi4:14b also showed significantly larger technical spillover, each with a median difference of 0.050. gemma3:4b showed the opposite pattern, with an argumentative distance exceeding the technical distance by a median of 0.025.

For qwen3:14b, the reported median difference in D^{unaff} rounded to 0.000, although the adjusted Wilcoxon result remained significant at $p_{\text{adj}} = 0.007$. This is not contradictory because the signed-rank test evaluates the ranks and directions of all nonzero paired differences rather than the rounded median alone. The distribution contained more negative than positive differences, indicating that argumentative unaffected-profile distances were generally larger, although the central difference rounded to zero at three decimal places.

gemma3:27b was the only model for which none of the three outcomes differed significantly between technical and argumentative responses. Its median differences were zero for both the total-score shift and affected-profile distance, while the small positive difference in unaffected-profile distance did not remain significant after correction. By contrast, phi4:14b showed significant response-type differences across all three outcomes: a larger technical score reduction, a larger argumentative affected-profile distance, and greater technical spillover into unaffected criteria.

Overall, the statistical comparisons confirm that rubric adaptation was strongly model dependent. Cross-model differences were especially pronounced for technical responses, where most model pairs differed in total-score change and both profile-distance measures.

Argumentative responses showed less separation in total-score shifts but retained substantial heterogeneity in the affected criterion profile. Response type did not exert a uniform effect across models: for some systems, technical responses produced larger score reductions or broader spillover, whereas others showed greater affected-criterion modification for argumentative responses. The observed behavior therefore cannot be reduced to a single model ranking or a general technical-versus-argumentative tendency.

3.7. Correspondence Between Model and Human Score Shifts

The differences between the two human evaluation conditions were compared with the score shifts produced by each model. For every response, the human cross-condition shift was calculated as the mean score assigned by E3 and E4 under Rubric B minus the mean score assigned by E1 and E2 under Rubric A. The corresponding model shift was calculated as the Rubric B score minus the Rubric A score. Their correspondence was evaluated using the mean absolute shift error, defined as the mean absolute difference between the model-generated shift and the human cross-condition shift for the same responses.

The human cross-condition shift was positive on average. The mean increase was 0.637 points overall, with a larger increase for technical responses, at 0.754 points, than for argumentative responses, at 0.519 points. The corresponding median shifts were 0.5 points overall, 1.0 point for technical responses, and 0.5 points for argumentative responses. Because Rubrics A and B were applied by different evaluator pairs, these cross-condition differences are descriptive and cannot be interpreted as a pure effect of the rubric weighting change.

The analysis included 1000 matched responses for each model, except phi4-mini:3.8b, for which one structurally invalid Rubric B output reduced the overall sample to 999 responses. The technical analyses included 500 responses for each model, except phi4-mini:3.8b, for which 499 responses were available. All argumentative analyses included 500 responses. No missing score was imputed. The 95% confidence intervals were estimated using 10,000 student-cluster bootstrap replicates.

Table 12 summarizes the overall correspondence between the model-generated and human cross-condition score shifts, including the mean shifts, their signed difference, the mean absolute shift error, and exact shift agreement.

Table 12. Overall correspondence between model and human cross-condition score shifts.

Model	Mean Model Shift	Mean Human Shift	Signed Difference	Mean Absolute Shift Error [95% CI]	Exact Shift Agreement (%) [95% CI]
qwen3:4b	−0.182	0.637	−0.819	1.498 [1.432, 1.565]	11.10 [9.10, 13.20]
qwen3:14b	−0.363	0.637	−1.000	1.521 [1.435, 1.606]	11.90 [9.80, 14.20]
gemma3:4b	−0.284	0.637	−0.921	1.605 [1.513, 1.694]	10.20 [8.20, 12.20]
gemma3:27b	−1.004	0.637	−1.641	1.942 [1.831, 2.050]	7.50 [5.70, 9.50]
llama3.1:8b	−1.239	0.637	−1.876	2.229 [2.121, 2.337]	6.50 [4.80, 8.30]
llama3.1:70b	−0.631	0.637	−1.268	1.686 [1.583, 1.788]	10.00 [8.10, 12.10]
phi4-mini:3.8b	−0.731	0.635	−1.366	1.749 [1.663, 1.839]	8.31 [6.61, 10.02]
phi4:14b	−1.001	0.637	−1.638	1.977 [1.866, 2.090]	8.00 [6.30, 9.80]

All eight models produced negative mean score shifts, whereas the mean human cross-condition shift was positive. Consequently, the signed difference, calculated as the mean model shift minus the mean human shift, was negative for every model. qwen3:4b showed the smallest directional discrepancy, with a mean model shift of −0.182 points and a signed difference of −0.819 points. The largest directional discrepancy occurred for llama3.1:8b, whose mean shift of −1.239 points produced a signed difference of −1.876 points.

The lowest overall mean absolute shift error was obtained by qwen3:4b, at 1.498 points, closely followed by qwen3:14b at 1.521 points and gemma3:4b at 1.605 points. These three models also produced the smallest negative mean total-score shifts. Their lower correspondence errors therefore resulted primarily from changes that remained closer to zero rather than from reproducing the positive average direction observed between the human evaluation conditions.

llama3.1:8b produced the largest overall mean absolute shift error, at 2.229 points. phi4:14b and gemma3:27b followed at 1.977 and 1.942 points, respectively. These models produced some of the largest negative total-score shifts and consequently remained farther from the predominantly positive differences observed between the human conditions.

Exact reproduction of the human cross-condition shift was uncommon for every model, although this measure was partly constrained by the different granularities of the model and human shifts. Exact shift agreement ranged from 6.50% for llama3.1:8b to 11.90% for qwen3:14b. Even the models with the lowest mean absolute shift errors therefore reproduced the human cross-condition difference exactly in only a limited proportion of responses.

The magnitude of the correspondence error also differed between technical and argumentative responses. Table 13 reports the mean absolute shift errors separately for the two response types.

Table 13. Mean absolute shift errors for technical and argumentative responses.

Model	Technical Shift Error [95% CI]	Argumentative Shift Error [95% CI]
qwen3:4b	1.472 [1.382, 1.564]	1.523 [1.416, 1.637]
qwen3:14b	1.564 [1.449, 1.683]	1.477 [1.358, 1.604]
gemma3:4b	1.586 [1.475, 1.697]	1.623 [1.494, 1.760]
gemma3:27b	2.024 [1.886, 2.163]	1.859 [1.728, 1.994]
llama3.1:8b	2.550 [2.403, 2.697]	1.907 [1.759, 2.059]
llama3.1:70b	1.846 [1.710, 1.983]	1.525 [1.399, 1.656]
phi4-mini:3.8b	1.864 [1.749, 1.983]	1.635 [1.517, 1.757]
phi4:14b	2.322 [2.161, 2.482]	1.631 [1.510, 1.755]

For technical responses, qwen3:4b showed the lowest mean absolute shift error, at 1.472 points. qwen3:14b and gemma3:4b followed at 1.564 and 1.586 points, respectively. The largest technical error occurred for llama3.1:8b, at 2.550 points, followed by phi4:14b at 2.322 points and gemma3:27b at 2.024 points. The particularly large error for llama3.1:8b reflects the contrast between the positive human technical shift of 0.754 points and the model's mean technical reduction of 1.594 points.

For argumentative responses, qwen3:14b produced the lowest mean absolute shift error, at 1.477 points. qwen3:4b and llama3.1:70b followed closely at 1.523 and 1.525 points, respectively. The largest argumentative error was obtained by llama3.1:8b, at 1.907 points, followed by gemma3:27b at 1.859 points.

The response-type pattern was not uniform across models. qwen3:4b and gemma3:4b showed slightly lower correspondence errors for technical responses. In contrast, qwen3:14b, gemma3:27b, llama3.1:8b, llama3.1:70b, phi4-mini:3.8b, and phi4:14b showed lower errors for argumentative responses. The largest technical–argumentative differences occurred for phi4:14b and llama3.1:8b, whose technical errors exceeded their argumentative errors by approximately 0.69 and 0.64 points, respectively.

Overall, the results indicate limited correspondence between the cross-rubric score shifts produced by the models and the differences observed between the two human evaluation conditions. Although the models differed in the magnitude of this discrepancy,

none reproduced the positive mean direction of the human cross-condition shift. Because the two rubric conditions were applied by different evaluator pairs, these comparisons remain descriptive and cannot be interpreted as agreement with a controlled estimate of the isolated rubric effect. Nevertheless, they demonstrate that rubric-matched model–human alignment within each condition does not necessarily imply similar adaptation when criterion weights change.

4. Discussion

The results obtained in this study reveal substantial differences in the scores assigned by local open-weight LLMs under alternative point allocations within the same analytic rubric structure. Although cross-rubric score differences were observed for all evaluated models, these differences varied in magnitude, direction, and localization across criteria and response types. These patterns have implications for rubric sensitivity, model-specific evaluative behavior, and the distinction between localized criterion adaptation and broader score redistribution.

4.1. Rubric Sensitivity and Criterion-Level Adaptation

The consistent reduction in mean total scores under Rubric B was observed across all evaluated models. Because the two rubrics retained the same four criteria, the same criterion definitions, and the same 10-point total, this result cannot be attributed to a smaller scoring range or to a general change in the assessment task. These findings are consistent with rubric sensitivity, but the same-rubric baseline indicates that run-to-run variability may also contribute to the observed score differences. This finding is consistent with previous evidence that LLM-based grading is sensitive to the structure and formulation of assessment instructions. Changes in prompting strategies, scoring information, and rubric refinement have been shown to modify automated grading outcomes, sometimes without producing consistent improvements in human–model agreement [27–30].

However, rubric sensitivity alone does not establish criterion-consistent adaptation. All models assigned lower total scores on average, but the alternative rubric did not prescribe a lower final grade. It reduced the maximum contribution of Clarity while increasing the maximum contribution of Completeness or Dialecticality by the same amount. A localized response would therefore involve a redistribution between the two affected dimensions, with limited changes elsewhere. The observed negative totals indicate that, for many models, the reduction in Clarity was not fully offset by gains in the criterion whose maximum increased.

The distinction between total-score stability and criterion-level stability is particularly important. For several models, an unchanged final grade frequently concealed modifications within the analytic profile. A final score alone therefore provides insufficient evidence about whether an LLM applied two rubrics identically or redistributed credit between their dimensions. This supports the broader argument that criterion-based grading should not be evaluated exclusively through agreement on aggregate scores, because different internal judgments can produce the same numerical total. The predominance of partially concordant rather than fully concordant changes also indicates that score shifts in the affected criteria were often asymmetric. A model frequently reduced Clarity while leaving Completeness or Dialecticality unchanged, or increased the gaining criterion while leaving Clarity unchanged. From a grading perspective, this describes partial directional concordance, but does not establish whether it resulted from rubric adaptation, run-to-run variability, or both.

The raw and normalized criterion shifts further clarify the differences between rubric conditions. A lower raw Clarity score did not always correspond to a stricter assessment

of clarity relative to the points available. Several models awarded fewer numerical points after the maximum was reduced from three to two while simultaneously awarding a larger proportion of that reduced maximum. Conversely, some criteria whose maxima increased received more raw points but a smaller proportion of the available three-point scale. These patterns indicate that the paired criterion scores were not related by simple proportional rescaling. Such behavior is compatible with evidence that LLM graders may apply implicit feature weights that differ across models and performance groups [34]. When the explicit maximum assigned to a criterion changes, the model may not merely convert an existing judgment to a new numerical range; it may also reinterpret how strongly the response satisfies the dimension. This could help explain why models with similar total-score shifts sometimes exhibited substantially different affected-profile distances and directional-concordance patterns. However, this mechanism remains hypothetical, and run-to-run variability may also contribute to the observed score differences.

The observed cross-rubric variation in Accuracy, Terminology, Coherence, and Originality shows that score changes were not confined to the reweighted criteria. Previous studies have reported criterion-dependent model–human agreement and dependencies among nominally separate assessment dimensions [19,24–26]. Criteria-entanglement bias has also been identified in multidimensional LLM evaluation, where judgments assigned to distinct dimensions become excessively dependent on one another [35]. The present findings are consistent with this concern. For some models, changing the point allocations of Clarity and Completeness or Dialecticality was accompanied by substantial revisions to criteria whose definitions and maxima were unchanged. The same-rubric baseline indicates that this broader score variation cannot be attributed exclusively to the alternative criterion allocation.

Taken together, these findings support a multidimensional interpretation of rubric fidelity. A rubric-sensitive grader should modify its scores when the assessment priorities change, but sensitivity is only one requirement. The change should also be directionally consistent with the revised allocation, concentrated in the affected criteria, and sufficiently stable in dimensions that remain formally unchanged. The evaluated models satisfied these requirements to different degrees. Consequently, successful adaptation cannot be inferred from the presence of a total-score shift alone; it requires simultaneous examination of the final grade, the affected criterion profile, compensating changes, and spillover into unrelated dimensions.

4.2. Human Reference Variability and Model–Human Alignment

The human evaluations showed that the reference scores were not fully invariant across evaluators. Both evaluator pairs achieved moderate-to-strong agreement, but systematic differences in scoring severity remained. Thus, stronger agreement in the ordering of responses did not eliminate differences in the absolute score levels assigned by the evaluators. For rubric-matched model–human alignment, the human scores were retained as separate comparison points rather than being reduced to a single averaged reference. For this alignment analysis, averaging the two evaluators within each rubric condition would have concealed the differences between them and produced a reference score that neither evaluator had actually assigned. Previous studies have likewise evaluated agreement, consistency, and scoring accuracy as distinct properties of LLM-based grading rather than treating them as interchangeable indicators [19,32].

The results further showed that similarity between mean scores was not sufficient to establish strong model–human alignment. A model may therefore show a small aggregate bias while still differing substantially from the human score on individual responses, because positive and negative errors can cancel when averaged. Exact agreement provided

an additional qualification. These results are compatible with the higher QWK values because ordinal agreement and exact score reproduction capture different aspects of model–human alignment.

The cross-condition analysis further distinguishes rubric-matched alignment from shift correspondence. A model may align comparatively well with human evaluators within each rubric while failing to reproduce the change observed between rubric conditions. Within-condition alignment measures proximity to a human reference under a fixed rubric, whereas shift correspondence measures whether the model modifies its judgment in a similar direction and magnitude when the rubric changes. The two properties are therefore related but not equivalent. The human cross-condition difference must nevertheless be interpreted cautiously. Rubric A was evaluated by E1 and E2, whereas Rubric B was evaluated by E3 and E4. The positive human shift may therefore reflect both the redistribution of criterion maxima and differences in scoring severity between the evaluator pairs. It cannot be interpreted as a controlled estimate of the isolated rubric effect. The comparison remains informative as a descriptive assessment of whether the models reproduced the empirical difference observed between the two human conditions.

The grading information differed between evaluator types. The LLMs received the reference answer, expected content elements, and banned misconceptions, whereas the human evaluators used only the question, student response, criterion definitions, and rubric-specific point allocations. Model–human agreement therefore combines differences in evaluative judgment with differences in the information available during grading. Under this design, the agreement measures describe correspondence between the two operational evaluation procedures rather than interchangeability between human and model evaluators under matched informational conditions.

Overall, model–human alignment was evaluator-dependent, metric-dependent, and condition-dependent. No model dominated simultaneously across both rubrics, both human references, both response types, and all agreement measures. Reliable evaluation of rubric-responsive LLM graders should therefore preserve separate human reference scores and combine directional bias, response-level error, ordinal agreement, exact agreement, and cross-condition shift correspondence rather than reducing performance to a single coefficient or averaged score.

4.3. Cross-Model and Response-Type Specificity

The adaptation patterns were strongly model-specific. Cross-model differences were significant for total-score shifts, affected-profile distances, and unaffected-profile distances in both response categories. However, the magnitude of these differences was consistently greater for technical responses. These differences were not explained by model size alone. Within the evaluated model set, increasing parameter count therefore did not consistently improve rubric adaptation or criterion-level stability.

The results also distinguish sensitivity from controlled adaptation. A model that changed its scores strongly was not necessarily more responsive in a desirable manner. This pattern indicates that the magnitude of a score change should not be interpreted independently of its location within the rubric structure. Similarly, an unchanged total score may conceal compensating changes between affected criteria. Such behavior would be missed by analyses restricted to final-score differences. It also shows that total-score stability can represent either genuine rubric insensitivity or successful compensation between criteria, which have different implications for rubric fidelity.

Response-type effects were similarly nonuniform. The influence of response type therefore depended on both the model and the adaptation outcome being examined. The technical–argumentative contrast may reflect differences in the structure of the two as-

sessment tasks. Although both rubric modifications transferred one point from Clarity to another criterion, Completeness and Dialecticality represent different evaluative demands. The models were therefore not responding merely to an identical numerical redistribution. They were also interpreting the semantic relation between the reduced and increased criteria within each response category. The stronger cross-model differences observed for technical responses suggest that models varied considerably in how they separated clarity from substantive coverage. This indicates that agreement at the total-score level does not imply agreement in the criterion-level mechanism used to obtain that score.

Overall, the findings do not support a single model hierarchy that applies across all adaptation outcomes. Models that performed comparatively well for total-score stability were not always the strongest in criterion-level redistribution, unaffected-criterion stability, or human-shift correspondence. Rubric-responsive grading should therefore be evaluated for each model and response type separately. Model selection based only on average score agreement or parameter scale would overlook substantial differences in how the grading decision is internally redistributed across rubric criteria.

4.4. Methodological Implications and Study Limitations

The distinction between raw and normalized criterion shifts was also methodologically important. Raw differences show how many points were gained or lost after the rubric modification, whereas normalized differences account for the changing criterion maxima. A criterion may receive a higher raw score under a larger maximum without receiving a larger proportion of the available points. Reporting only one representation can therefore produce an incomplete interpretation of whether a model adapted to the revised weighting. The study further shows that stability should not be treated as a uniformly desirable outcome. Stability in unaffected criteria supports localization of the rubric response, whereas stability in the affected criteria may indicate rubric insensitivity. Similarly, an unchanged final score may represent either no adaptation or compensating changes between the criteria whose weights changed. The meaning of score stability depends on where it occurs within the rubric structure.

Human comparison introduces an additional methodological requirement. Model–human alignment under each rubric and correspondence between cross-rubric shifts assess different questions. A model may approximate human scores within both conditions while changing its evaluations differently when the rubric changes. Both forms of comparison are necessary when the objective is to determine whether an automated grader is not only accurate under a fixed rubric but also responsive to modifications in the assessment priorities.

Several limitations constrain the interpretation of the findings. First, the two human rubric conditions were evaluated by different pairs of evaluators. This design reduced potential carry-over and anchoring between rubric conditions but did not permit within-evaluator estimation of human rubric adaptation. In contrast, each LLM evaluated every response under both rubrics in a paired design. The observed human cross-condition shift therefore combines potential rubric effects with pair-specific differences in scoring severity. It provides a descriptive reference for the empirical difference between conditions but does not isolate the causal effect of the rubric modification.

Second, the rubric modifications were limited to a one-point redistribution between two existing criteria while the total score and remaining criterion definitions were preserved. This controlled design made criterion-level changes directly comparable, but it does not represent all forms of rubric revision. The dataset contained two response categories and a fixed set of questions. The technical and argumentative tasks provided distinct criterion structures, but the results may not generalize to other domains, educational levels, response lengths, or assessment formats. The study evaluated the models without rubric-specific

calibration or task-specific fine-tuning. The findings therefore describe the behavior of general-purpose instruction-tuned models rather than the upper performance limit of systems optimized for grading.

Third, the analysis focused on eight local open-weight models evaluated under a fixed inference configuration. Same-rubric repetitions revealed run-to-run variability despite unchanged prompts and generation parameters, so cross-rubric differences cannot be attributed exclusively to rubric redistribution. Only one repetition per condition was performed, without examining alternative prompt formulations or different decoding settings. The observed adaptation patterns may partly depend on the specific inference protocol.

Future work should use within-evaluator crossover designs with counterbalanced rubric order and an appropriate washout period, examine broader rubric transformations across additional domains and assessment formats, increase same-rubric repetitions under controlled prompt and decoding variations, and test rubric-specific calibration or task-specific fine-tuning while preserving criterion-level observability.

5. Conclusions

This study examined whether local open-weight LLMs adapt their grading behavior when the point allocation of an analytic rubric changes while the evaluated responses and overall score scale remain constant. The models generally reduced their total scores under Rubric B, although the human cross-condition scores increased on average. These shifts were not uniformly aligned with the intended redistribution between the affected criteria. In several cases, score changes extended to criteria whose definitions and maxima had remained unchanged, but the same-rubric baseline showed that these differences cannot be attributed exclusively to rubric redistribution. The analysis also showed that final-score stability can conceal meaningful internal redistribution. Evaluating only final scores would therefore have obscured both compensating adaptation and unintended spillover.

Overall, the findings indicate that rubric-responsive LLM grading should be assessed as a multidimensional process. Reliable evaluation requires separate examination of total-score changes, affected-criterion redistribution, unaffected-criterion stability, response-type dependence, evaluator-specific alignment, and correspondence with human cross-condition differences. A model should not be considered rubric-adaptive merely because its final score changes after the rubric is modified. Human comparisons further demonstrated that alignment under a fixed rubric and adaptation across rubric conditions are distinct properties.

Because different evaluator pairs applied the two rubrics, this cross-condition comparison remains descriptive and cannot isolate the rubric effect from differences in evaluator severity. Future work should extend this design to larger rubric transformations, additional domains, repeated scoring by the same human evaluators under counterbalanced conditions, and controlled variations in prompts and decoding settings. Such studies would help determine whether criterion-level adaptation can be improved through calibration or task-specific training and whether the observed patterns remain stable across assessment contexts.

Author Contributions: Conceptualization, C.A., A.A.A., M.V.C., A.C., S.M., M.V. and C.S.; methodology, C.A., A.A.A., M.V.C., A.C. and M.V.; software, C.A., A.A.A. and M.V.; validation, C.A., A.A.A., M.V.C., A.C., S.M., M.V. and C.S.; formal analysis, C.A., A.A.A., M.V.C. and M.V.; data curation, C.A., A.A.A., M.V.C., A.C. and M.V.; writing—original draft preparation, C.A., A.A.A., M.V.C. and M.V.; writing—review and editing, C.A., A.A.A., M.V.C., A.C., S.M., M.V. and C.S.; visualization, C.A., A.A.A. and M.V.; supervision, C.A., M.V.C., A.C., S.M. and C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study because it was a secondary methodological analysis that reused the same anonymous question texts and student responses from the GradeAgentOps dataset and did not involve any new participant recruitment or data collection.

Informed Consent Statement: Informed consent was waived because this study was a secondary methodological analysis based on existing anonymous data.

Data Availability Statement: The inference script used to construct the complete grading prompts for Rubrics A and B, the input dataset, and the experimental configuration are available at: https://github.com/anghelcata/RubricAdapt-LLM. (accessed on 26 August 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, D.-W.; Boey, M.; Tan, Y.Y.; Jia, A.H.S. Evaluating large language models for criterion-based grading from agreement to consistency. *NPJ Sci. Learn.* **2024**, *9*, 79. [CrossRef]
2. Brookhart, S.M. Appropriate Criteria: Key to Effective Rubrics. *Front. Educ.* **2018**, *3*, 22. [CrossRef]
3. Hu, X.; Gao, M.; Hu, S.; Zhang, Y.; Chen, Y.; Xu, T.; Wan, X. Are LLM-based Evaluators Confusing NLG Quality Criteria? In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 11–16 August 2024; pp. 9530–9570.
4. Ramesh, D.; Sanampudi, S.K. An Automated Essay Scoring Systems: A Systematic Literature Review. *Artif. Intell. Rev.* **2022**, *55*, 2495–2527. [CrossRef] [PubMed]
5. Misgna, H.; On, B.-W.; Lee, I.; Choi, G.S. A Survey on Deep Learning-Based Automated Essay Scoring and Feedback Generation. *Artif. Intell. Rev.* **2025**, *58*, 36. [CrossRef]
6. Emirtekin, E. Large Language Model-Powered Automated Assessment: A Systematic Review. *Appl. Sci.* **2025**, *15*, 5683. [CrossRef]
7. Grévisse, C. LLM-based automatic short answer grading in undergraduate medical education. *BMC Med. Educ.* **2024**, *24*, 1060. [CrossRef] [PubMed]
8. Pecuchova, J.; Benko, Ľ.; Drlik, M. Automated Grading of Open-Ended Questions in Higher Education Using GenAI Models. *Int. J. Artif. Intell. Educ.* **2025**, *35*, 3813–3846. [CrossRef]
9. Poličar, P.G.; Špendl, M.; Curk, T.; Zupan, B. Automated assignment grading with large language models: Insights from a bioinformatics course. *Bioinformatics* **2025**, *41*, i21–i29. [CrossRef] [PubMed]
10. Chen, Z.; Wan, T. Grading Explanations of Problem-Solving Process and Generating Feedback Using Large Language Models at Human-Level Accuracy. *Phys. Rev. Phys. Educ. Res.* **2025**, *21*, 010126. [CrossRef]
11. Mendonça, P.C.; Quintal, F.; Mendonça, F. Evaluating LLMs for Automated Scoring in Formative Assessments. *Appl. Sci.* **2025**, *15*, 2787. [CrossRef]
12. Li, Y.; Raković, M.; Srivastava, N.; Li, X.; Guan, Q.; Gašević, D.; Chen, G. Can AI support human grading? Examining machine attention and confidence in short answer scoring. *Comput. Educ.* **2025**, *228*, 105244. [CrossRef]
13. Wang, Y.; Huang, J.; Du, L.; Guo, Y.; Liu, Y.; Wang, R. Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. *Comput. Educ. Artif. Intell.* **2025**, *9*, 100481. [CrossRef]
14. Rodrigues, L.; Xavier, C.; Costa, N.; Gašević, D.; Mello, R.F. Is GPT-4 Fair? An Empirical Analysis in Automatic Short Answer Grading. *Comput. Educ. Artif. Intell.* **2025**, *8*, 100428. [CrossRef]
15. Panadero, E.; Jonsson, A.; Pinedo, L.; Fernández-Castilla, B. Effects of Rubrics on Academic Performance, Self-Regulated Learning, and self-Efficacy: A Meta-analytic Review. *Educ. Psychol. Rev.* **2023**, *35*, 113. [CrossRef]
16. Martin, P.P.; Kranz, D.; Graulich, N. Revealing Rubric Relations: Investigating the Interdependence of a Research Informed and a Machine Learning Based Rubric in Assessing Student Reasoning in Chemistry. *Int. J. Artif. Intell. Educ.* **2025**, *35*, 1465–1503. [CrossRef]
17. Hashemi, H.; Eisner, J.; Rosset, C.; Van Durme, B.; Kedzie, C. LLM-RUBRIC: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Bangkok, Thailand, 11–16 August 2024; pp. 13806–13834.
18. Yavuz, F.; Çelik, Ö.; Yavaş Çelik, G. Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *Br. J. Educ. Technol.* **2025**, *56*, 150–166. [CrossRef]
19. Seo, H.; Hwang, T.; Jung, J.; Kang, H.; Namgoong, H.; Lee, Y.; Jung, S. Large Language Models as Evaluators in Education: Verification of Feedback Consistency and Accuracy. *Appl. Sci.* **2025**, *15*, 671. [CrossRef]

20. Anghel, C.; Craciun, M.V.; Pecheanu, E.; Cocu, A.; Anghel, A.A.; Iacobescu, P.; Maier, C.; Andrei, C.A.; Scheau, C.; Dragosloveanu, S. CourseEvalAI: Rubric-Guided Framework for Transparent and Consistent Evaluation of Large Language Models. *Computers* **2025**, *14*, 431. [CrossRef]
21. Anghel, C.; Anghel, A.A.; Pecheanu, E.; Craciun, M.V.; Cocu, A.; Niculita, C. PEARL: A Rubric-Driven Multi-Metric Framework for LLM Evaluation. *Information* **2025**, *16*, 926. [CrossRef]
22. Fan, Z.; Wang, W.; Wu, X.; Zhang, D. SedarEval: Automated Evaluation using Self-Adaptive Rubrics. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024; pp. 16916–16930.
23. Li, Q.; Cui, L.; Kong, L.; Bi, W. Exploring the Reliability of Large Language Models as Customized Evaluators for Diverse NLP Tasks. In Proceedings of the 31st International Conference on Computational Linguistics (COLING), Abu Dhabi, United Arab Emirates, 19–24 January 2025; pp. 10325–10344.
24. Chu, S.; Kim, J.W.; Wong, B.; Yi, M.Y. Rationale Behind Essay Scores: Enhancing S-LLM's Multi-Trait Essay Scoring with Rationale Generated by LLMs. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, NM, USA, 29 April–4 May 2025; pp. 5811–5829.
25. Seßler, K.; Fürstenberg, M.; Bühler, B.; Kasneci, E. Can AI Grade Your Essays? A Comparative Analysis of Large Language Models and Teacher Ratings in Multidimensional Essay Scoring. In Proceedings of the 15th International Learning Analytics and Knowledge Conference, Dublin, Ireland, 3–7 March 2025; pp. 462–472.
26. Li, S.; Ng, V. Graph-Based Multi-Trait Essay Scoring. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 4–9 November 2025; pp. 33325–33351.
27. Lee, G.-G.; Latif, E.; Wu, X.; Liu, N.; Zhai, X. Applying Large Language Models and Chain-of-Thought for Automatic Scoring. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100213. [CrossRef]
28. Ariely, M.; Salman, A.; Yarden, A.; Alexandron, G. Reflective Prompt Engineering: A New Strategy for Automated Short Answer Scoring in Biology. *Int. J. Sci. Educ.* **2025**, 1–23. [CrossRef]
29. Xue, M.; Liu, Y.; Xiao, X.; Wilson, M. Automatic Prompt Engineering for Automatic Scoring. *J. Educ. Meas.* **2025**, *62*, 559–587. [CrossRef]
30. Norgaila, E.; Daniela, L.; Kalniņa, D. Reflective Prompt Engineering for Assessment Rubric Optimization: An Empirical Study of Human–AI Alignment. *Technol. Knowl. Learn.* **2026**, *31*, 1023–1038. [CrossRef]
31. Korthals, L.; Akrong, E.; Geller, G.; Rosenbusch, H.; Grasman, R.; Visser, I. Towards Reliable LLM Grading Through Self-Consistency and Selective Human Review: Higher Accuracy, Less Work. *Mach. Learn. Knowl. Extr.* **2026**, *8*, 74. [CrossRef]
32. Xue, M.; Xiao, X.; Liu, Y.; Wilson, M. On the Consistency of Automatic Scoring with Large Language Models. *Educ. Psychol. Meas.* **2026**, *86*, 861–882. [CrossRef] [PubMed]
33. Jukiewicz, M. A Systematic Comparison of Large Language Models for Automated Assignment Assessment in Programming Education: Exploring the Importance of Architecture and Vendor. *Comput. Educ. Open* **2026**, *10*, 100364. [CrossRef]
34. Wang, M.; Chen, Y.; Huang, X.; Lai, Y. Opening the Blackbox of LLM-Based Automated Essay Scoring: Insights into Feature Weighting Patterns and Score Validity. *Comput. Educ. Artif. Intell.* **2026**, *10*, 100568. [CrossRef]
35. Li, W.; Wang, X.; Yuan, S.; Xu, R.; Chen, J.; Dong, Q.; Xiao, Y.; Yang, D. Curse of Knowledge: Your Guidance and Provided Knowledge Are Biasing LLM Judges in Complex Evaluation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, 4–9 November 2025; pp. 14900–14924.
36. Koo, R.; Lee, M.; Raheja, V.; Park, J.I.; Kim, Z.M.; Kang, D. Benchmarking Cognitive Biases in Large Language Models as Evaluators. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 517–545.
37. Anghel, C.; Anghel, A.A.; Craciun, M.V.; Cocu, A.; Vulpe, D.-E.; Andrei, C.A.; Maier, C.; Scheau, C.; Dragosloveanu, S.; Cergan, R. GradeAgentOps: A Verification-First Framework for Evidence-Anchored LLM Exam Grading. *AI* **2026**, *7*, 198. [CrossRef]
38. Alibaba Cloud. Qwen/Qwen3-4B Model Card. Available online: https://huggingface.co/Qwen/Qwen3-4B (accessed on 24 June 2026).
39. Alibaba Cloud. Qwen/Qwen3-14B Model Card. Available online: https://huggingface.co/Qwen/Qwen3-14B (accessed on 8 August 2026).
40. Google. google/gemma-3-4b-it Model Card. Available online: https://huggingface.co/google/gemma-3-4b-it (accessed on 24 June 2026).
41. Google. google/gemma-3-27b-it Model Card. Available online: https://huggingface.co/google/gemma-3-27b-it (accessed on 24 June 2026).
42. Meta. meta-llama/Llama-3.1-8B-Instruct Model Card. Available online: https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct (accessed on 8 August 2026).
43. Meta. meta-llama/Llama-3.1-70B-Instruct Model Card. Available online: https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct (accessed on 24 June 2026).

44. Microsoft. Phi-4-Mini-Instruct Model Card. Available online: https://huggingface.co/microsoft/Phi-4-mini-instruct (accessed on 19 June 2026).
45. Microsoft. Phi-4 Model Card. Available online: https://huggingface.co/microsoft/phi-4 (accessed on 8 August 2026).
46. Ollama Inc. Ollama API Documentation. Available online: https://docs.ollama.com/api/introduction (accessed on 8 August 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.