JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Conditional Returns to Language Adaptation: An Adaptation Ladder Study of Indonesian Short Answer Grading

Imam Rangga Bakti [1,2,*], Handaru Jati [3], and Nurkhamid [3]

[1] Doctoral Program of Engineering Science, Faculty of Engineering, Universitas Negeri Yogyakarta, Yogyakarta 55281, Indonesia; e-mail: imamrangga.2023@student.uny.ac.id
[2] Department of Informatics Engineering, Faculty of Computer Science, Universitas Pasir Pengaraian, Rokan Hulu 28558, Indonesia; e-mail: imamranggabakti@gmail.com
[3] Department of Electronics and Informatics Engineering Education, Faculty of Engineering, Universitas Negeri Yogyakarta, Karangmalang, Sleman 55281, Yogyakarta, Indonesia; e-mail: handaru@uny.ac.id; nurkhamid@uny.ac.id
* Corresponding Author : Imam Rangga Bakti

Abstract: Indonesian has several adapted language models for automatic short answer grading. Studies disagree about whether specialization is worthwhile; each test one point on the adaptation spectrum, so a losing language-specific model is indistinguishable from useless specialization. An ordinal comparison replaces it: a three-rung ladder of released checkpoints, from no Indonesian adaptation through regional to national continued pretraining, in two families whose rungs share a base model and parameter count. In the Gemma family, the first two rungs are siblings rather than ancestor and descendant, which the analysis carries through. Across 30,960 grading calls on 36 independent questions, with the question as the unit of inference, returns are conditional rather than cumulative. Regional adaptation helps both families, by 0.029 and 0.018 pooled quadratic weighted kappa (QWK). The second step adds little on the weaker base, 0.016, and harms the stronger one, −0.029. That decline is established in one family and directional in the other. Labelled examples drawn from the question being graded halve the remaining advantage, and examples from other questions attenuate it further, from +0.069 to +0.005, a difference of −0.064 [−0.107, −0.022], although they grade less accurately in absolute terms. Spreading the demonstrated score levels reduces it to +0.029. All six models moved in the direction predicted from their initial scoring bias at four shots, and five of six at eight shots. Error decomposition locates the gain in a shift in grading severity, not sharper discrimination, and the shift runs in opposite directions in the two families. A 124-million-parameter Indonesian encoder, included as a deployment reference, reaches 0.825 QWK in the social science domain at 596 MB, within 0.027 of the best 9-billion-parameter configurations. Language adaptation is therefore a conditional recalibration. Its value depends on the base model, on labeled answers already held, and on which score levels they display.

Keywords: Automatic short answer grading; Continued pretraining; Educational assessment; In-context learning; Indonesian natural language processing; Language adaptation; Large language models; Low-resource languages.

Received: August, 10th 2026
Revised: October, 13th 2026
Accepted: October, 14th 2026
Published: Novemver, 1st 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) licenses (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

Short-answer questions require students to construct responses in their own words while keeping them brief enough to be graded at scale. This combination makes grading costly: every response must be read, and the same standard must be applied consistently across hundreds of responses. Automatic short-answer grading (ASAG) addresses this challenge. The corpus most frequently used for Indonesian ASAG contains responses from 534 respondents scored by seven subject experts [1]. Subsequent work has reported lower error rates [2] using separately assembled corpora. Consequently, published results are based on different evaluation subsets and are not directly comparable.

Large language models offer a potential solution, and adapting them for languages with fewer resources than English has become an active research direction. Indonesian is relatively well served, with resources ranging from a mature language-specific encoder [3] and a regional language model family [4] to instruction-tuned Indonesian generative models [5] and general-purpose multilingual models [6], [7]. The practical question is which point on this spectrum should be deployed. However, the literature offers inconsistent findings. Li and Liu [8] and Schneider et al. [9] found that general-purpose models could match or outperform language-specific alternatives, whereas Padó et al. [10] developed a German-specific Transformer to address the persistent resource limitations of non-English models. These studies address a related model-selection decision, yet none cites the others.

A shared limitation is that each study evaluates a single point on the adaptation spectrum. When a language-specific model underperforms, such a design cannot distinguish between two explanations: language specialization may offer limited benefits, or the particular model evaluated may simply be weaker. In this sense, the comparison is binary, whereas adaptation is an ordinal process. The benefits of adaptation need not increase with every additional stage. If returns are bounded, a further stage may even reduce performance, making the point at which additional adaptation ceases to help a practically important question. Investigating this issue requires comparing adaptation stages that share a common base model rather than comparing independently developed models. Sections 2.2 and 2.4 discuss the relevant adaptation strategies and explain why previous studies cannot resolve this question.

Indonesian provides an opportunity to conduct such a comparison. The regional model family [4] and the Indonesian continued-pretraining family [11], [12] are built on publicly available base checkpoints, allowing adaptation stages to be compared at the same parameter count within each family and the pattern to be examined across two model families. This study makes three contributions:

- An ordinal evaluation of language adaptation. It replaces binary comparisons between general-purpose and language-specific models with an adaptation ladder of publicly released checkpoints, using the question as the unit of inference.
- An analysis of the relationship between adaptation and labelled examples. It examines whether labelled examples can substitute for further adaptation and whether the source of those examples affects this relationship.
- An error-mechanism and deployment analysis. It decomposes grading errors to investigate whether adaptation changes grading severity or discrimination, and includes a fine-tuned Indonesian encoder as a deployment reference. The analysis distinguishes effects supported by the results from those that remain directional.

The remainder of this paper is organized as follows. Section 2 reviews related work on Indonesian ASAG, language adaptation, grading behaviour, and the research gap. Section 3 describes the adaptation ladder, dataset preparation, experimental protocol, evaluation measures, statistical analysis, and cost measurement. Section 4 presents and discusses the results, including adaptation effects, error decomposition, robustness analyses, deployment costs, and follow-up experiments. Section 5 concludes the paper and outlines directions for future research.

2. Related Work

2.1. Automatic short answer grading for Indonesian

Automatic short answer grading predates the transformer era by two decades. Burrows et al. [13] trace its development from concept mapping and information extraction to corpus-based and machine-learning approaches, identifying the recurring obstacle that a correct answer may share almost no surface form with the reference. Mohler and Mihalcea [14] established the semantic-similarity formulation that still dominates the field, and the shared task of Dzikovska et al. [15] fixed a common protocol for cross-system comparison. Essay scoring followed a parallel path, neural models displacing feature engineering from the mid-2010s [16]. Indonesian ASAG research has largely followed the semantic similarity paradigm. Haidir and Purwarianti [1] built the corpus used here and applied contextual word embeddings with linear regression, reporting error on the Indonesian portion substantially worse than on the English portion they used for comparison. Salim et al. [2] achieved lower error on separate

corpora they assembled themselves, and excluded responses scored zero on the ground that a model need not learn from incorrect answers.

Short answer grading with large language models has since become an active line of its own. The closest published analogue is Chang and Ginter [17], who graded Finnish responses with GPT-4 and obtained a quadratic weighted kappa above 0.6 in 44 per cent of one-shot settings, a lower-resource language graded by a general-purpose model with agreement that is useful but short of human. Closest in spirit within Indonesian is Darnoto and Firmawan [18], who exploit linguistic proximity between Indonesian regional languages as a transfer signal. Their scope differs: transfer between regional languages rather than the number of adaptation stages applied to one language. Among the Indonesian ASAG studies reviewed here, none compares adaptation strategies while holding the base model constant. The corpus scored by the largest panel of experts offers the most reliable reference labels available for the task, which makes it the appropriate testbed.

2.2. Adaptation strategies for languages with fewer resources

Work on lower-resource languages divides into three strategies that are rarely evaluated against one another. The first builds language-specific models from scratch: IndoBERT [3] for Indonesian, and the German model of Padó et al. [10], both justified by resource scarcity in non-English settings.

The second performs continued pretraining on a general-purpose base, a principle established by Gururangan et al. [19]: a second stage of pretraining on in-domain or in-task text improves downstream performance over the original model, and the gain depends on how far the target distribution lies from the pretraining distribution. Language adaptation is the same operation with language as the target distribution. Regional variants such as SEA-LION [4] adapt to a language group, while the Indonesian continued-pretraining family used here [20] targets one national language together with two of its regional languages. Two properties matter: the gain is defined relative to the original model, so it is not independent of how good that model already is, and nothing guarantees that a second adaptation stage adds to a first. Neither has been tested for language adaptation.

The third relies on general-purpose multilingual models without adaptation: Schneider et al. [9] found multilingual encoders matching systems specialized by question type, subject, and language, and Li and Liu [8] found GPT-4 outperforming a Japanese-specific model. Kortemeyer [21] reached the intermediate position on standard benchmarks, finding a general-purpose GPT-4 comparable to hand-engineered systems but below models that had received specialized training, which is what an ordering rather than a dichotomy would predict.

2.3. What automated graders actually respond to

A parallel line of work questions whether automated scorers respond to content at all. Kumar et al. [22] found BERT-based essay scorers both overstable, with scores barely moving when content changed, and oversensitive, with scores shifting when a few words were added. Filighera et al. [23] showed the analogous vulnerability in short answer grading, where inserting adjectives and adverbs can mislead a grader. Rodrigues et al. [24] found GPT-4 more generous than human markers, and Grévisse [25] found GPT-4 markedly stricter than teachers on a task where Gemini 1.0 Pro was not, so two models given one rubric can depart from the human standard in opposite directions. Emirtekin and Özarslan [26] report agreement between 0.585 and 0.640 quadratic weighted kappa that declines as the cognitive demand of the item rises. Aggregate accuracy can conceal what a grader is doing, and a grader's departure from the human standard is often a matter of severity rather than of discrimination, which motivates the error decomposition in Section 4.4 and the severity profile in Section 4.7. Misgna et al. [27] survey the area and identify non-English evaluation as an open direction.

2.4. Research Gap

Five studies frame the design choices adopted here. Weegar and Idestam-Almquist [28] reported quadratic weighted kappa between 0.49 and 0.57 with classical models, and Jati et al. [29] established that deployability deserves reporting alongside accuracy; this study reports cost as its deployability measure because the protocol here is deterministic. Generative work sets a more demanding band, 0.677 with GPT-4 across ten ASAP short-answer questions [30] and 0.70 against 0.75 between human markers [31], so the distance worth reporting is the

distance to human agreement rather than to zero. Cong et al. [32] find fine-tuned encoders still strongest on this task while rubric-based prompting of open-weight models can match them without annotated data, which frames the encoder result of Section 4.6.

The gap is therefore specific. The contradiction between [8], [9], and [10] concerns how much language adaptation is worthwhile. Yet none of the contributing studies varied the number of adaptation stages applied, and none cites the methodological result [19] that defines the gain from continued pretraining relative to the model being adapted. Resolving it requires an ordinal design in which released adaptation stages share a base model, with the checkpoint lineage of each family stated.

3. Proposed Method

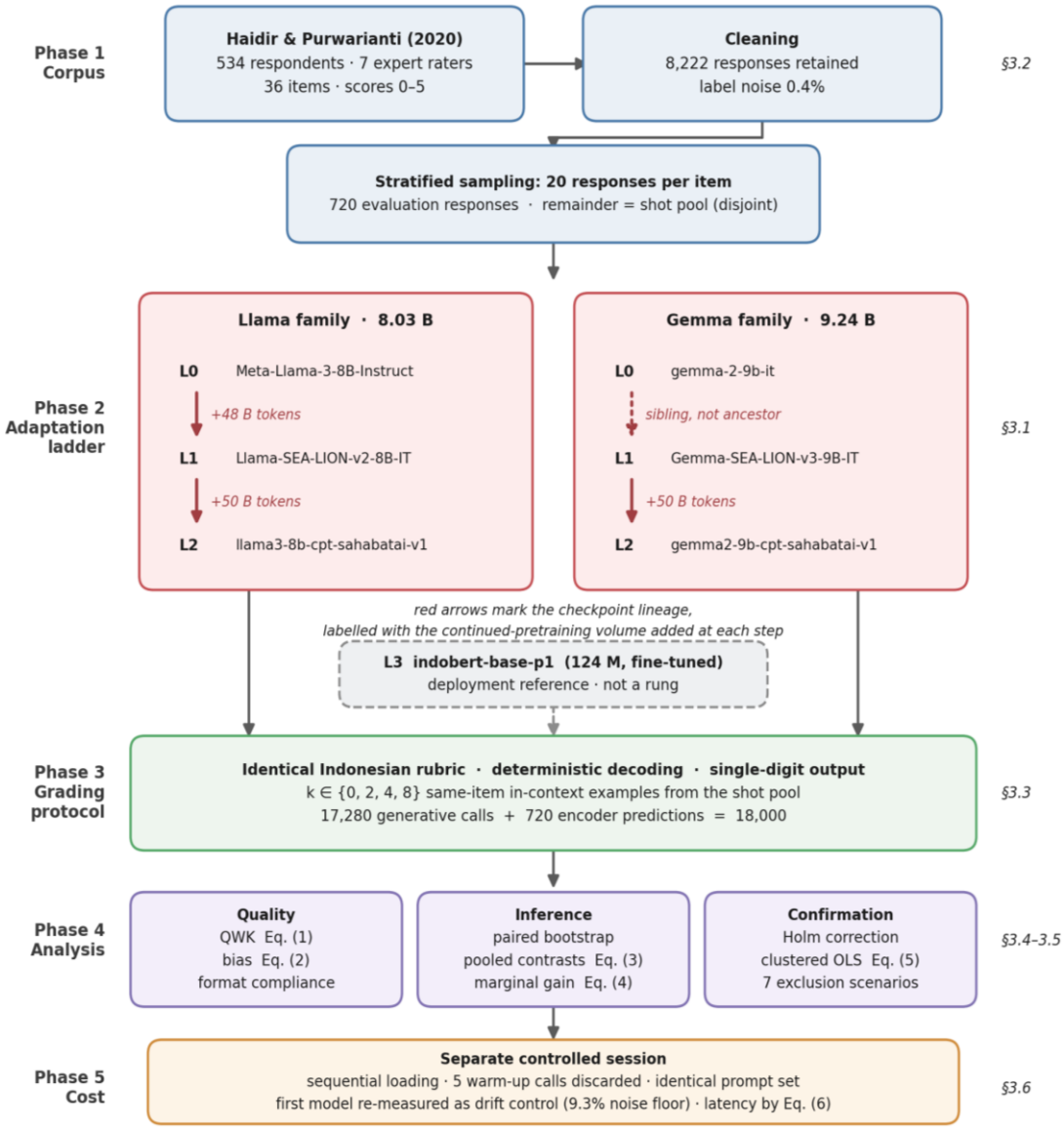


Figure 1. Experimental pipeline

3.1. The adaptation ladder and its control

The independent variable is the adaptation stage of a publicly released checkpoint, treated as an ordinal factor with three rungs, shown in Table 1. The rungs are ordered by the released adaptation stage rather than by any single quantity, such as token volume, because both the volume and language composition of continued pretraining change between stages. Level L0 is a general-purpose multilingual model with no Indonesian adaptation, L1 adds continued pretraining on Southeast Asian languages, and L2 adds continued pretraining

targeting Indonesian together with Javanese and Sundanese. A fourth model, an Indonesian-specific encoder denoted L3, is included as a deployment reference rather than a rung. Because it differs from the generative models in architecture, parameter count, and mode of use, it is excluded from every adaptation contrast.

Table 1. The adaptation ladder and the deployment reference

Level	Adaptation	Params	Cumulative CPT tokens	Llama family	Gemma family
L0	None (general-purpose)	8.03 B/9.24 B	none	Meta-Llama-3-8B-Instruct	gemma-2-9b-it
L1	Regional (Southeast Asia)	8.03 B/9.24 B	48 B (Llama) / 200 B (Gemma)	Llama-SEA-LION-v2-8B-IT	Gemma-SEA-LION-v3-9B-IT
L2	National (Indonesian CPT)	8.03 B/9.24 B	98 B (Llama) / 250 B (Gemma)	llama3-8b-cpt-sahabatai-v1	gemma2-9b-cpt-sahabatai-v1
L3 (reference, not a rung)	Indonesian-specific encoder	0.12 B	not applicable	indobert-base-p1 (fine-tuned)	n/a

Within each family, the three rungs share the same base model and an identical parameter count: 8.03 billion in the Llama family and 9.24 billion in the Gemma family, although their lineage relationships differ. The comparison holds model size and architecture fixed while examining performance across the adaptation stages represented by the released checkpoints. Running the ladder in two families therefore replicates the stage comparison rather than repeating one controlled intervention.

Two choices constrain the design. In the Llama family, the regional rung uses the second-generation regional model because the third-generation model is built on Llama 3.1, whereas the Indonesian continued-pretraining model is built on Llama 3; using the newer regional model would reintroduce a base-model difference. L3 remains outside the ladder because no Indonesian-specific encoder of comparable size is available, and it can be fine-tuned but not evaluated through the same prompting procedure.

The configuration is identical across the six generative checkpoints: Q4_K_M quantisations of the repositories listed in Table 1, served through llama.cpp with every model layer offloaded to the GPU. The context length is 4,096 tokens, temperature is set to 0, top-p to 1, and the output is capped at six tokens. A fixed seed of 42 is used for decoding and sampling. No system role is used; instead, the rubric is prepended to each model's first user turn. The experiments use a single NVIDIA Tesla T4 GPU with 16 GB of memory. L3 is fine-tuned at full precision for three epochs, with a batch size of 16, a learning rate of 2×10^{-5}, and a maximum sequence length of 256 tokens. Appendix A provides the file digests and remaining specifications, while Section 4.11 discusses the consequences of quantisation.

The training provenance of these checkpoints further clarifies what each adaptation stage represents. In the Llama family, L1 was continued from the unadapted base on 48 billion tokens across five Southeast Asian languages. L2 was continued from that checkpoint on a further 50 billion tokens targeting Indonesian, Javanese, and Sundanese, giving 98 billion cumulative tokens. In the Gemma family, the corresponding figures are 200 billion tokens across eleven Southeast Asian languages, followed by a further 50 billion tokens, giving 250 billion cumulative tokens. All figures are as declared by the publishers.

Thus, the changes between adaptation stages involve not only training-data volume but also language composition, which narrows from a regional group to Indonesian and two regional languages. These factors are confounded by construction because no publicly available checkpoint family varies one while holding the other constant. The ladder therefore establishes the effects of successive adaptation stages as distributed by the publishers, rather than the effect of adaptation volume in isolation. The publishers do not disclose the per-source composition of their pretraining mixtures, data-filtering criteria, or complete training hyperparameters. Consequently, the conclusions concern the publicly released checkpoints as distributed rather than a controlled adaptation procedure.

3.2. Corpus, evaluation set, and shot pool

The corpus is the Indonesian short answer collection of Haidir and Purwarianti [1]. It holds responses from 534 respondents to Biology items (*saintek*) and Geography items (*soshum*), each scored 0 to 5 by seven subject experts, across 36 items, 15 training and 3 held-out per domain. The source repository carries no tagged release, and its data directory has since been rewritten, so every figure reported here reads one pinned commit, 90ed4c13, recorded in the Data Availability Statement. Of the 9,086 rows in that commit, 61 carry a blank response, none carry a score outside 0 to 5, and 803 duplicate an earlier row exactly on item, response, and score, leaving 8,222 responses. Label noise is low: 30 pairs of identical responses to the same item carry more than one score, about 0.4 per cent.

The corpus contains thousands of responses but only 36 independent questions, and it is the question count that bounds any claim about generalization to new questions. One property of the released files requires care: item numbering in the held-out files restarts at 1, so a naive merge on item number pairs responses with the wrong reference answer and raises no error at all. Merging is therefore performed per split with a compound identifier. Grading through prompting requires no training data, so evaluating only the held-out items would discard 30 of the 36 items and leave too few clustering levels to estimate between-item variance. All 36 items are therefore evaluated, with the held-out items additionally analyzed as an unseen-prompt condition. For each item, 20 responses are sampled with stratification across the six score levels, giving 360 responses per domain and 720 per experimental condition. The remainder forms the shot pool, disjoint from the evaluation set and drawn from the same item, so in-context examples introduce no leakage. Responses average 10.1 whitespace-delimited words and 2.44 score points in the science domain, and 12.2 words and 2.29 in the social science domain.

3.3. Grading protocol

A single Indonesian rubric describing the six score levels is held identical across all models and conditions, since prompts tuned per model would no longer attribute score differences to the rung. Each call presents the item stem, the expert reference answer, and one student response, and instructs the model to reply with a single digit from 0 to 5. Decoding is deterministic at temperature zero with output capped at six tokens. Because one family's chat template rejects a system role, the rubric is prepended to the first user turn for all models.

The number of in-context examples, k, takes the values 0, 2, 4, and 8, following Brown et al. [33] and spanning the range in which demonstrations are known to act. Min et al. [34] found that demonstrations mainly convey the label space rather than input-label mappings, which predicts diminishing returns beyond a few examples and motivates testing substitution against language adaptation.

Examples are drawn from the shot pool of the same item, one response per score level in ascending order until k is reached, so below k=6 the demonstrations occupy the lower part of the scale. Each item's shot pool fixes which levels the rule displays, so coverage is reported that way. At k=2 the rule shows scores 0 and 1 for 34 of the 36 items, and at k=4 scores 0 to 3 for 27 of the 36, a higher level standing in wherever a lower one is absent. At k=8, it shows all six levels for 20 of the 36. Even coverage below k=6 is arithmetically impossible. Two consequences follow. The demonstrations are identical for all six models within a cell, so coverage is a constant across rungs and cannot contribute to any contrast between them. And the number of examples is not varied independently of the scores displayed, which Section 4.5 tests with a condition spreading the levels across the scale at k=4. Same-item sampling guarantees that no evaluated response appears among its own demonstrations, and conveys how the rubric applies to that question rather than the label space alone, which Section 4.10 separates.

Seven of the 720 evaluated responses have a text-identical counterpart in the shot pool carrying a different score. No item-response-score triple is shared between the two sets, so this is the corpus label noise of Section 3.2 appearing at evaluation-set level rather than leakage, and under few-shot it can place a contradictory demonstration in one per cent of prompts. Section 4.11 carries it forward. Outputs are parsed by accepting a bare digit, and cases where a digit must be recovered from longer text, or where none appears, are recorded so that format compliance is reported rather than silently absorbed.

Level L3 is fine-tuned rather than prompted. The encoder receives the item stem paired with reference answer and student response, trains for three epochs on the shot pool of the corresponding domain, and is evaluated on the same evaluation set.

3.4. Measures

Grading quality is reported as quadratic weighted kappa [35], appropriate because the scale is ordinal. For K levels, with observed frequencies O and expected frequencies E from the marginals over score levels i and j, the coefficient (QWK, Eq. (1)) is

$$\kappa = 1 - \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij}\ O_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij}\ E_{ij}}, \qquad w_{ij} = \frac{(i-j)^2}{(K-1)^2} \tag{1}$$

where w_{ij} penalizes disagreement quadratically, so a two-level disagreement counts four times a one-level one. Here, K=6. The coefficient is computed overall, and per item, the latter averaged so no single item dominates.

Exact agreement, agreement within one level, and mean absolute error are reported alongside it. Kappa corrects for chance using the marginals, so a change in how predictions are distributed moves the coefficient even when the ordering of responses is unchanged, which bears on this study directly since Section 4.4 locates the effect of adaptation in exactly such a shift. The rung ordering is identical under all four measures: mean absolute error falls from 0.908 to 0.826 across the Llama ladder and rises from 0.749 to 0.785 across the Gemma one, so no conclusion depends on the choice of coefficient. Mean bias, which carries the mechanism analysis in Section 4.4, is the signed counterpart of absolute error over N responses:

$$\text{bias} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i) \tag{2}$$

Here $\hat{y}$ denotes the model score and y the expert score of a response. Positive values denote leniency, negative values severity. Reporting bias separately is necessary because a grader can reduce error at one end of the scale while increasing it at the other, which is what Section 4.4 finds. Format compliance is the proportion of outputs that are a bare digit.

3.5. Statistical analysis

The independent unit throughout is the question, not the response. The main grid issues 18,000 grading calls clustered within 36 items, so resampling responses would treat 360 answers per cell as independent and understate every interval. As a sensitivity analysis, the contrasts are also estimated with the response as the unit, and the resulting change in standard errors is reported in Section 4.1. Contrasts between rungs are therefore estimated by a cluster bootstrap [36] over items, 10,000 replications, resampling items with replacement within each domain and recomputing quadratic weighted kappa from the summed per-item confusion matrices. Holm correction is applied across the 48 per-cell contrasts. Per-cell contrasts rest on 18 items each, so contrasts are additionally pooled across the eight cells of each family-contrast combination by inverse-variance weighting, which yields the primary inferential estimates. Those eight cells share the same items, so one item draw is shared across all of them within a replication, and the dependence between conditions enters the pooled interval. The weights are fixed at their point estimates and do not vary between replications. Where Section 4.10 compares a demonstration rule with its matched condition, both carry the weights estimated on the main grid, so the difference between two rules is a weighted average of per-cell differences rather than a comparison of two differently weighted averages. Given per-cell estimates Δ_c with cluster-bootstrap standard errors s_c over C cells,

$$\bar{\Delta} = \frac{\sum_{c=1}^{C} w_c\,\Delta_c}{\sum_{c=1}^{C} w_c}, \qquad SE(\bar{\Delta}) = \left(\sum_{c=1}^{C} w_c\right)^{-1/2}, \qquad w_c = \frac{1}{s_c^2} \tag{3}$$

so that cells estimated more precisely contribute more. Because the eight cells share items, the standard error above is the analytic benchmark under independence, and the reported

intervals are cluster-bootstrap intervals. Where an interval spans zero, its bound is reported as the largest effect the data rule out, so null results carry information.

Because the ladder is ordinal, the quantity of interest is not only whether a rung differs from the unadapted base but how much each successive step adds:

$$\delta_1 = \kappa_{L1} - \kappa_{L0}, \qquad \delta_2 = \kappa_{L2} - \kappa_{L1}, \qquad \text{decline} = \delta_1 - \delta_2 \tag{4}$$

each estimated by the same bootstrap and pooled by Eq. (3), with δ_1 and δ_2 taken from the same item draws so that their difference inherits the correct covariance. A positive decline indicates that the return to a further stage is bounded. A negative δ_2 indicates a harmful second step.

Effects on absolute error are estimated by ordinary least squares with standard errors clustered by item [37]. Writing L for rung, F for family, S for subject domain, and k for number of examples,

$$|\hat{y}_{ic} - y_{ic}| = \beta_0 + \beta_L L + \beta_F F + \beta_S S + \beta_k k + \gamma_{LF}(L \times F) + \gamma_{Lk}(L \times k) + \gamma_{LS}(L \times S) + \varepsilon_{ic} \tag{5}$$

where i indexes responses and q items, and negative coefficients denote smaller error. Here F indicates the Gemma family, S the subject domain, and L and k are dummy-coded against $\mathrm{L0}$ and $\mathrm{k} = 0$. A mixed-effects model with item as a random intercept was attempted and abandoned: the item variance estimate converged to the parameter boundary, and subject domain is nested completely within item. Robustness is assessed across seven exclusion scenarios: full data, responses scored zero removed following [2], scores zero and one removed, training items only, held-out items only, zero-shot only, and few-shot only. Kappa magnitudes are not comparable across scenarios that alter the label range, so only direction and significance are interpreted there.

3.6. Cost measurement

Latency recorded during the grading runs proved uninterpretable: one model appeared 45 per cent slower than its two same-sized counterparts, and because it was loaded first after a session restart, the difference was confounded with run order. Cost is therefore measured in a dedicated session under four controls: all models loaded sequentially on the same device, five warm-up calls per model discarded, an identical prompt set for every model, and the first model measured again at the end as a drift control. The two measurements of that model differed by 9.3 per cent, treated as the noise floor, so differences smaller than about ten per cent are not interpreted anywhere in this article. The anomaly resolved under this protocol. The unadapted Gemma model emits three output tokens where every other model emits one, and the extra tokens account for the difference at a decode rate consistent with the hardware. Latency is normalized to a single output token by Eq. (6), writing t_m for mean measured latency, n_m for mean output tokens, and estimating the per-token decode cost τ from a same-family pair of models m' and m'' differing mainly in output length:

$$t_m^* = t_m - (n_m - 1)\,\tau, \qquad \tau = \frac{t_{m'} - t_{m''}}{n_{m'} - n_{m''}} \tag{6}$$

The failure of the unadapted model to stop after producing the required digit is noted here as a finding about output-constraint compliance rather than about speed.

4. Results and Discussion

The main grid comprises 17,280 generative grading calls and 720 encoder predictions, and the two follow-up conditions of Sections 4.5 and 4.10 add 12,960 generative calls, 30,960 in all, each on the same 36 independent questions. Of the 30,240 generative outputs, 13 required a digit to be recovered from longer text, six of them in the main grid and all 13 from the regionally adapted Gemma model. None failed to parse, so no result below is affected by parsing loss.

4.1. Where adaptation pays off

Table 2 reports the grid. Quality lies between 0.592 and 0.825 when averaged across the two domains, and between 0.537 and 0.852 for individual domain cells. For context, classical methods on this task report 0.49 to 0.57 [28], and contemporary generative work on English

corpora reports 0.677 [30] and 0.70 to 0.75 between human markers [31], neither comparable term by term, so the baselines of Section 4.2 are the only directly comparable figures. More examples improve quality almost everywhere. The spread between rungs is far wider in the Llama family, whose unadapted model starts at 0.592, than in the Gemma family, which starts at 0.731. Of the 48 per-cell contrasts, 8 survive Holm correction under the item cluster bootstrap, uniformly positive in the Llama family and uniformly negative in the Gemma family. The largest is +0.143, for national adaptation over the unadapted base in the science domain at zero-shot. Each per-cell contrast rests on 18 items and is correspondingly underpowered, so the pooled estimates in Table 3 carry the primary inference.

Table 2. Quadratic weighted kappa by family, adaptation level, and number of examples

Family	Level	k=0	k=2	k=4	k=8	Fine-tuned
gemma	L0	0.731	0.768	0.759	0.805	n/a
gemma	L1	0.748	0.796	0.775	0.825	n/a
gemma	L2	0.729	0.780	0.728	0.787	n/a
llama	L0	0.592	0.706	0.697	0.725	n/a
llama	L1	0.701	0.708	0.741	0.741	n/a
llama	L2	0.719	0.733	0.759	0.780	n/a
indobert	L3	n/a	n/a	n/a	n/a	0.763

Table 3. Pooled contrasts by family

Family	Contrast	Pooled ΔQWK	CI low	CI high	p	Effect
llama	L1−L0	+0.0285	+0.0081	+0.0475	0.006	gain
llama	L2−L0	+0.0616	+0.0396	+0.0823	<0.001	gain
llama	L2−L1	+0.0163	+0.0020	+0.0293	0.025	gain
gemma	L1−L0	+0.0175	+0.0054	+0.0299	0.003	gain
gemma	L2−L0	−0.0166	−0.0309	−0.0021	0.026	loss
gemma	L2−L1	−0.0292	−0.0376	−0.0207	<0.001	loss

Applying Eq. (3), regional adaptation is an advantage in both families, +0.029 in the Llama family and +0.018 in the Gemma family, both with intervals excluding zero. National adaptation behaves differently. In the Llama family, it adds further value, reaching +0.062 over the unadapted base and +0.016 over the regional rung, the latter with an interval whose lower bound is +0.002. In the Gemma family, the contrast against the regional rung is −0.029, also with an interval excluding zero. All six pooled contrasts retain their sign and their significance when the question rather than the response is the independent unit. Moving between units inflates the standard error by a median factor of 1.16, ranging from 0.77 to 2.05 across the 48 cells.

Table 4. Ordinary least squares (OLS) regression on absolute scoring error, Eq. (5). Reference categories: Llama, L0, k = 0

Term	Coef.	Cluster SE	p
Family (Gemma)	−0.1590	0.0417	<0.001
k = 2	−0.1042	0.0278	<0.001
k = 4	−0.0819	0.0397	0.039
k = 8	−0.1708	0.0446	<0.001
Rung L1	−0.0215	0.0524	0.681
Rung L2	−0.0990	0.0666	0.138
Rung L1 × Gemma	+0.0059	0.0366	0.872

Expressed as marginal gains, Eq. (4), the second step does not exceed the first in either family, although in the Llama family this holds for the point estimates only. In the Llama family δ_1 = +0.029 [+0.008, +0.048] and δ_2 = +0.016 [+0.002, +0.029]. In the Gemma family δ_1 = +0.018 [+0.005, +0.030] and δ_2 = −0.029 [−0.038, −0.021]. The decline is

established in the Gemma family, $\delta_1 - \delta_2$ = +0.047 [+0.031, +0.062], p < 0.001, and is directional but not significant in the Llama family, +0.012 [−0.012, +0.034], p = 0.30. There, the data rule out a decline larger than about 0.034 and an increase larger than about 0.012. In the Gemma family, the second step reverses, and only there do the data separate the decline from zero. Section 4.7 takes up what the pattern means. A complementary clustered regression on absolute error, Eq. (5), finds the same pattern in Table 4: national adaptation interacts with family at +0.119, p = 0.008, where a positive coefficient denotes greater error. The interactions of rung with number of examples and with subject domain are not significant.

4.2. Baselines under matched supervision

Reported figures for this corpus in earlier work were obtained on different evaluation subsets and are not comparable with ours. Three classical baselines were therefore evaluated on the same 720 responses and the same 36 items: a per-item mean score as a floor, TF-IDF character n-gram cosine similarity with linear regression, and IndoBERT mean-pooled embeddings with linear regression and no fine-tuning. Table 5 reports them. Each item receives its own regression, fitted on labelled examples drawn from that item's shot pool, with the count capped at 2, 4, and 8 to match the few-shot conditions and then left uncapped at roughly 208 per item.

Table 5. Classical baselines under matched supervision

Method	Labelled examples per item	QWK *saintek*	QWK *soshum*	QWK mean
B1 TF-IDF + linear regression	2	0.184	0.098	0.141
B1 TF-IDF + linear regression	4	0.456	0.417	0.437
B1 TF-IDF + linear regression	8	0.507	0.583	0.545
B1 TF-IDF + linear regression	≈208 (all)	0.523	0.668	0.596
B2 IndoBERT + linear regression	2	0.184	0.098	0.141
B2 IndoBERT + linear regression	4	0.492	0.443	0.467
B2 IndoBERT + linear regression	8	0.572	0.564	0.568
B2 IndoBERT + linear regression	≈208 (all)	0.646	0.668	0.657

The per-item mean floor is 0.000 by construction on the held-out items and is omitted from the table. At every matched level of supervision, every generative configuration in Table 2 exceeds the strongest baseline: by at least 0.565 QWK at two examples per item, 0.229 at four, and 0.157 at eight. The uncapped baseline reaches 0.657 using roughly twenty-five times the supervision of the k = 8 condition, and five of the six generative configurations exceed that figure with no labelled examples at all. The exception is the unadapted Llama model at 0.592. The same encoder used in two ways is informative. Frozen IndoBERT embeddings with a fitted regression reach 0.657, whereas the same encoder fine-tuned end-to-end reaches 0.763, a difference of 0.106 attributable to fine-tuning rather than to representation.

A second protocol tests the baselines where they are strongest. It trains on all labelled responses of the 30 training items, roughly 2,884 in the science domain and 3,400 in the social science domain, and tests on the held-out items alone. Frozen IndoBERT embeddings with a fitted regression then reach 0.517 quadratic weighted kappa in the science domain and 0.804 in the social science domain, and TF-IDF reaches 0.397 and 0.769. The social science figure is the important one: with roughly three thousand labelled answers, a classical method matches or exceeds every zero-shot generative configuration in that domain. The generative advantage documented above is therefore specific to the low-supervision regime, and in the social science domain it disappears once several thousand answers from other questions of the domain have been labelled. These baselines calibrate the generative results on this protocol. No comparison against contemporary generative systems on a shared benchmark is attempted.

4.3. Labelled examples substitute for adaptation on a weak base

Testing whether the advantage shrinks as examples are supplied requires estimating on the claimed quantity rather than a proxy. On quadratic weighted kappa, the pooled Llama

contrast for national adaptation over the unadapted base is +0.116 [+0.061, +0.172] at zero-shot and +0.055 [+0.032, +0.075] with examples, a difference of +0.061 [+0.009, +0.118], p = 0.022, with both quantities and their difference from the same item draws. In the Gemma family, the same comparison is not significant, p = 0.55, so substitution appears only on a weak basis. The effect survives the move to item-level inference but is weaker than a response-level analysis would suggest, and its magnitude rather than its existence is what remains uncertain.

Min et al. [34] found that generic demonstrations mainly convey the label space and the shape of the input, whereas the demonstrations here are drawn from the question being graded, so the question is whether the substitution depends on the source of the examples or only on their presence. Section 4.10 settles it with demonstrations drawn from the other questions of the domain at a matched number of examples; the substitution does not require examples drawn from the same question, and examples from other questions attenuate the advantage further rather than restoring it. On the weaker base, labelled examples therefore substitute in part for the benefit of language adaptation, within the task and the two domains evaluated, and the attenuation is not confined to examples drawn from the question being graded. Nothing here establishes that the same substitution holds across other tasks, other domains, or other few-shot protocols. Examples from other questions also grade less accurately in absolute terms, and the comparison runs between zero and at most eight examples on the 36 questions of this corpus.

4.4. What adaptation actually changes

Aggregate quality conceals the mechanism. Table 6 decomposes the change in absolute error by reference score, and Fig. 2 shows the bias curves.

Table 6. Error decomposition by reference score, Llama L0 to L2. Positive values denote reduced error

Reference score	n	Total change in error	95% CI	p
0	520	+184	[+107, +302]	<0.001
1	568	+200	[+116, +294]	<0.001
2	444	+1	[−56, +57]	0.977
3	476	−91	[−145, −33]	0.004
4	432	−113	[−183, −50]	<0.001
5	440	+57	[+12, +102]	0.013
Net	2,880	+238	[+28, +481]	0.022

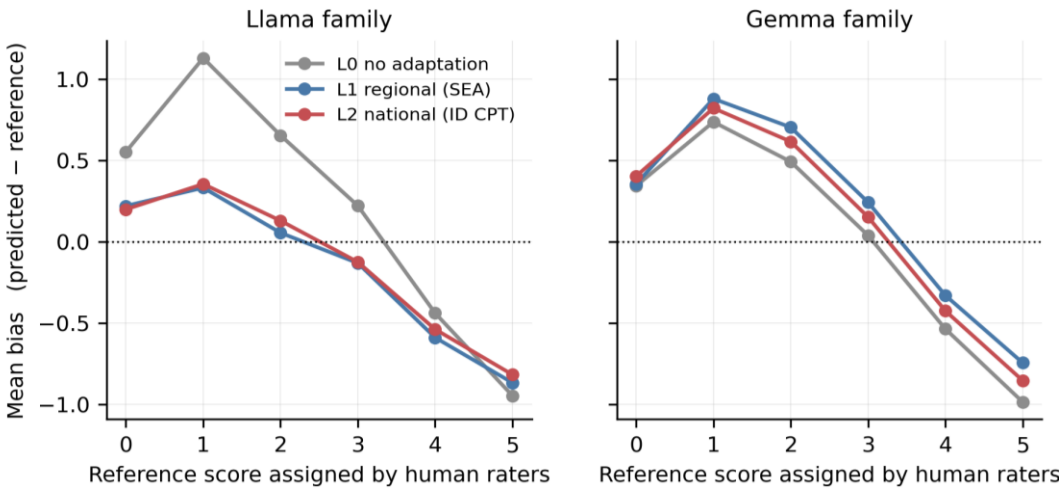


Figure 2. Mean bias by human reference score

Adaptation is not a uniform improvement. Error falls on responses experts scored 0 or 1 and rises on those scored 3 or 4, with a net change of +238 [+28, +481], positive only because errors at the bottom were much larger to begin with. Overall bias, Eq. (2), moves

from +0.249 for the unadapted model to −0.101 for the adapted one, a shift of about 0.35 points towards severity. On responses deserving zero, the unadapted model awards 0.552 and over-scores 26.0 per cent, against 0.198 and 9.6 per cent for the adapted model. On responses deserving one, it awards 2.129 against 1.354, with over-scoring falling from 66.5 to 34.5 per cent. Both comparisons are paired on identical responses.

The pattern is therefore a change in grading severity rather than a general improvement in reading Indonesian answers: it helps where the grader was too lenient and hurts where it was already strict enough. The direction of this change originates in the base model rather than in adaptation. In the Llama family, mean bias across all conditions shifts from +0.249 at L0 to −0.130 at L1 and −0.101 at L2; in the Gemma family it moves from +0.058 to +0.222 and then to +0.160. Gemma without adaptation looks calibrated on average yet remains strict at the top band, awarding 0.537 points too few to responses worth 4 and 0.986 too few to those worth 5, and the first adaptation step makes it more generous at every level. The beneficial Gemma step gains at the opposite end of the scale: error falls by 39 points at score 4 and 106 points at score 5, while rising by 55 and 48 points at scores 1 and 2. The second step damages the upper half again, adding 31, 51 and 48 points of error at scores 3, 4 and 5.

4.5. Robustness

In Table 7, the Llama effect survives all seven scenarios, including the held-out subset, where it reaches +0.066 [+0.028, +0.103] on six questions alone. Six questions constrain generalization to unseen prompts only weakly, but they do establish that the effect is not an artefact of the thirty items available while the protocol was being designed. The effect also survives removal of all responses scored zero, following [2], because its largest component lies at reference score one rather than zero. The Gemma effect survives three of seven, which is why the claim there is stated as national adaptation not exceeding regional adaptation rather than harming performance relative to no adaptation. The contrast against the regional rung survives six of the seven, failing only under zero-shot prompting, where it is −0.017 [−0.035, +0.001]. The contrast against the unadapted base survives three.

Table 7. Pooled L2−L0 contrast under seven exclusion scenarios

Exclusion scenario	Llama ΔQWK	95% CI	Sig.	Gemma ΔQWK	95% CI	Sig.
Full data (0–5)	+0.062	[+0.040, +0.082]	yes	−0.017	[−0.031, −0.002]	yes
Zero scores removed	+0.128	[+0.091, +0.163]	yes	−0.021	[−0.046, +0.004]	no
Scores 0 and 1 removed	+0.108	[+0.055, +0.163]	yes	−0.018	[−0.052, +0.016]	no
Training items only (30)	+0.061	[+0.036, +0.085]	yes	−0.014	[−0.031, +0.003]	no
Held-out items only (6)	+0.066	[+0.028, +0.103]	yes	−0.030	[−0.053, −0.007]	yes
Zero-shot only	+0.116	[+0.061, +0.172]	yes	−0.012	[−0.033, +0.010]	no
Few-shot only	+0.055	[+0.032, +0.075]	yes	−0.019	[−0.035, −0.002]	yes

A second selection rule tests the same contrast. Below k=6, the rule of Section 3.3 shows only the lower part of the scale, so the condition reported in Table 10 repeats k = 4 with the four levels spread across the range, at 0, 2, 3 and 5. Spreading them raises absolute quality in all six cells, by +0.003 to +0.053, and preserves the rung ordering in both families, so the ladder of Section 4.1 does not rest on the restricted coverage of the original rule. The size of the contrast does rest on it, in the family where adaptation had an advantage to lose. The Llama advantage falls from +0.069 [+0.042, +0.094] to +0.029 [+0.004, +0.055], a difference of −0.040 [−0.068, −0.010], p = 0.008. In the Gemma family, the deficit is no longer distinguishable from zero under the spread rule, −0.009 [−0.050, +0.039] against −0.047 [−0.074, −0.021], though the change between the rules is directional rather than established, +0.037 [−0.007, +0.088], p = 0.11. The effect of the number of examples is therefore not separable from the score levels those examples display. Effects are broadly distributed rather than driven by a few items: 27 of 36 items improve in the Llama family, median +0.032, against 13 of 36 in the Gemma family, median −0.017. All 20 Gemma contrasts that do not survive the correction lie within ±0.05 of zero, the largest at 0.044. An advantage of one tenth of a kappa point is therefore ruled out rather than merely undetected.

4.6. Cost and deployability

Table 8 summarizes the resource requirements of the evaluated models, including file size, VRAM usage, inference latency, loading time, and best QWK. Figure 3 illustrates the relationship between the best grading quality achieved by each model and its GPU memory requirements.

Table 8. Resource requirements

Model	File size (GB)	VRAM (MB)	ms / answer	Load time (s)	Best QWK
llama/L0	4.921	5250	163.2	20.72	0.773
llama/L1	4.921	5252	168.6	20.03	0.815
llama/L2	4.921	5252	172.0	19.16	0.817
gemma/L0	5.761	7366	179.9	24.40	0.837
gemma/L1	5.761	7366	179.9	25.03	0.852
gemma/L2	5.761	7366	179.8	25.15	0.827
indobert/L3	0.498	596	9.1	9.52	0.825

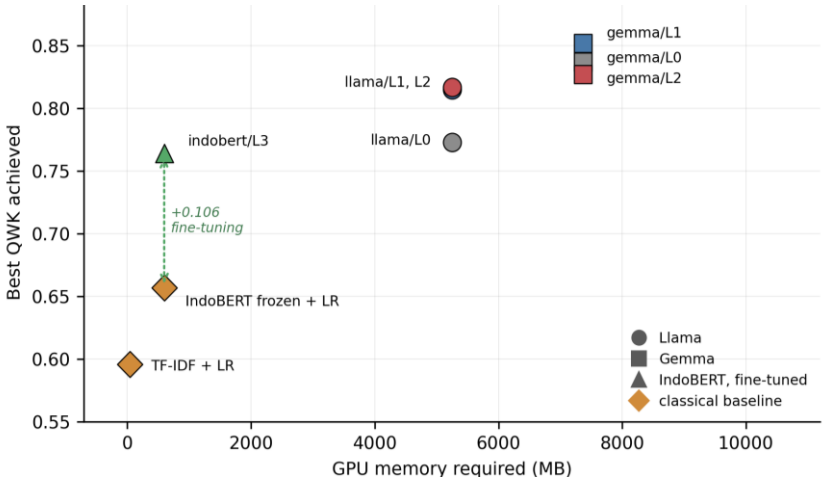


Figure 3. Best grading quality against GPU memory

Within each family, memory and latency are largely determined by the base model rather than the adaptation rung. As shown in Table 8 and Figure 3, the Llama variants occupy approximately 5,250 MB of VRAM, while the Gemma variants require about 7,366 MB. Latency differences fall below the 9.3% noise floor after normalization by Eq. (6) and are therefore not interpreted. The fine-tuned encoder is the outlier, with 124 million parameters, 596 MB of memory usage, and 9.1 ms per answer. As shown in Table 8, it requires roughly one-eleventh of the VRAM and one-twentieth of the normalized inference latency of the generative models, while achieving 0.702 QWK in the science domain and 0.825 in the social science domain. In the social science domain, the three best generative configurations outperform the encoder, whose QWK is 0.027 below the best result of 0.852. In the science domain, the encoder's QWK is 0.100 below the best generative result.

Supervision is not matched: the encoder is fine-tuned on approximately 208 labelled answers per item, whereas the generative models use at most eight examples. Therefore, these results do not establish parity in grading quality under equivalent supervision. Nevertheless, the results indicate that this level of grading quality can be achieved with substantially lower hardware requirements when labelled answers are already available, consistent with the conditions favouring fine-tuned encoders discussed by Cong et al.[32]. All evaluated models fit on a single free-tier accelerator, but only the encoder can run without one.

4.7. Resolving the contradiction, and why the return is bounded

An ordinal view of adaptation reconciles Li and Liu [8], Schneider et al. [9], and Padó et al. [10] without requiring any of them to be wrong about the point they measured. The general model wins in [8] in a comparison that omits the regional rung, where the benefit sits here,

and on a base strong enough to leave little for national adaptation to add. The parity reported in [9] is what a bounded return predicts. The case for a language-specific model in [10] is consistent with a weak available base. These are readings the ordinal account makes available, not demonstrated causes: the three studies use different corpora, models, and protocols, and Section 4.1 states that figures across publications are not comparable term by term.

The severity profile of Section 4.4 suggests why. A general improvement in comprehension predicts a benefit spread across the score range and accumulating across stages. The data fit a narrower description: a change in where on the scale the grader is calibrated, and the benefit of that change is large only while there is miscalibration left to remove. A bounded return follows from that description, and it is consistent with how the gain from continued pretraining is defined in [19], relative to the model being adapted rather than in absolute terms.

The two families differ. The unadapted Llama model is lenient throughout, and adaptation moves it towards strictness: the first stage removes most of that leniency, the second has less to remove. The unadapted Gemma model is near calibration on average and strict at the top of the scale. Adaptation moves it towards generosity: the first stage corrects the top, the second does not repeat that correction, since mean bias falls back from +0.222 to +0.160, yet it worsens scores 3 to 5. The bound, then, is on the miscalibration available for adaptation to correct, a property of the base model rather than of the procedure. Section 2.3 sets out independent support for severity as the main axis of disagreement, where two models given one rubric depart from the human standard in opposite directions, and agreement falls as cognitive demand rises. That is the shape of the mid-range cost reported here. If a grader calibrates largely on surface fluency, as Kumar et al. [22] report, more Indonesian text would shift the threshold at which fluent but empty responses receive credit, without changing what the grader attends to. That step is an interpretation. This article establishes the recalibration, its direction in each family, and its cost in the middle of the scale, not the representational change that produces them.

4.8. Practical guidance

The findings translate into conditional recommendations rather than a ranking, collected in Table 9.

Table 9. Recommendation matrix

Condition	Recommendation	Evidence
Weak base, no labelled examples	Adapt to national rung (L2)	ΔQWK +0.116 at k=0, CI [+0.061, +0.172]
Weak base, labelled answers to the same questions	Adaptation helps but matters less	ΔQWK +0.055, CI [+0.032, +0.075]; difference from k=0 +0.061, p = 0.022
Strong base	Stop at regional rung (L1)	L1−L0 +0.018 [+0.005, +0.030]; L2−L1 −0.029 [−0.038, −0.021]
Any base	Regional adaptation is safe	L1−L0 positive in both families, $p \leq 0.006$
Graded answers to the same questions already held	Supply examples from those questions	Main effect of k: −0.082 to −0.171 absolute error, $p \leq 0.039$
Separating non-answers from answers	Adaptation helps substantially (weak base)	Error at reference score 0−1 falls by 384 on the weak base, $p < 0.001$
Discriminating among mid-range answers	Adaptation is counterproductive	Error at reference score 3−4 rises by 204 on the weak base, $p < 0.001$
Constrained hardware or on-premises data, labels in hand	Use the L3 encoder, given labelled answers per item	596 MB, 9.1 ms/answer, 0.825 QWK in *soshum*, 0.027 below the best 9B configuration

Adaptation is not a general improvement to be applied wherever affordable: on a strong base it moved quality the wrong way relative to the regional rung. And it is not additive with labelled examples, a point that carries a scope condition the table cannot hold, since it rests on demonstrations drawn from the same 36 questions being graded and on a comparison between zero and at most eight such demonstrations. This finding implies that an institution already holding graded answers to its own questions gains less from a further adaptation stage than one starting without any, and gains most where those answers belong to the questions it wants graded. It is not a general rule for deploying short answer grading, and in particular it

does not establish that assembling examples is a better use of effort than adaptation for questions that have never been graded. Which examples are supplied matters separately from whether any are; examples from other questions attenuate the advantage further, but they also grade less accurately in absolute terms, and absolute grading quality is the reason for preferring same-question examples. Section 4.10 reports the figures.

4.9. Findings that did not fit the expectation

Two results ran against expectations formed before the data were seen, and a third test was abandoned. Benefit was expected to be larger in the social science domain, on the reasoning that Geography answers depend more on everyday Indonesian than Biology answers depend on internationally shared vocabulary. No interaction between rung and domain reached significance in either family, though with two domains, absence of interaction cannot be distinguished from insufficient power. Removing responses scored zero, as practiced in [2], was expected to eliminate the effect and did not, because its largest component lies at reference score one. An attempt to test whether the advantage concentrates on responses containing non-standard Indonesian was abandoned on measurement grounds. An indicator built from a published colloquial Indonesian lexicon and a dictionary word list captured scientific terminology and affixed standard forms rather than non-standardness, and restricting it to the lexicon alone left only 82 of 720 responses with any colloquial token, a minimum detectable effect more than twice the family effect. The phenomenon is rare in formal examination responses.

4.10. Demonstrations drawn from other questions

The claim of Section 4.3 turns on the few-shot design, not on the adaptation ladder. Demonstrations in the main grid come from the question being graded, so they supply question-specific supervision together with the label space, and Section 4.3 could not separate the two. The follow-up draws them from the pooled shot responses of the other 17 items of the domain, at $k = 4$, where the per-cell estimates are most precise, and at $k = 8$, where same-item demonstrations gain most. It holds the rest fixed: evaluation set, rubric, decoding settings, evaluated responses, all six checkpoints across both domains, and the number of demonstrations and their spread across score levels. Only the source question changes.

Table 10 sets this follow-up against its matched cells, alongside the selection-rule condition of Section 4.5. In the Llama family, the advantage of the national rung over the unadapted base is +0.069 [+0.042, +0.094] with same-item demonstrations at k=4 and +0.005 [−0.032, +0.043] with cross-item demonstrations, a difference of −0.064 [−0.107, −0.022], $p = 0.002$. At $k = 8$, the pair is +0.051 [+0.004, +0.095] and −0.014 [−0.071, +0.047], a difference of −0.065 [−0.140, +0.008], $p = 0.087$. Cross-item demonstrations therefore do not return the advantage towards its zero-shot value of +0.116 [+0.061, +0.172] but attenuate further what the same-item condition had left of it, to an estimate whose interval includes zero. Within the task and the two domains evaluated, the substitution of Section 4.3 therefore extends to demonstrations drawn from other questions of the same domain. The Gemma family does not move with it, and the asymmetry is part of the result. There the national rung never leads the unadapted base, starting at −0.012 [−0.033, +0.010] with no examples, and cross-item demonstrations leave that gap where it is at k=4, −0.047 [−0.074, −0.021] against −0.028 [−0.058, +0.002], a difference of +0.019 [−0.015, +0.053], $p = 0.27$. At k=8 the gap is larger under cross-item demonstrations, −0.052 [−0.085, −0.018] against −0.018 [−0.041, +0.004], though the difference between the rules is directional, −0.034 [−0.070, +0.004], $p = 0.086$, so the dependence is established only in the family where adaptation had something to add.

Attenuation of the contrast is not equivalence of grading quality. Cross-item demonstrations grade worse in all six Llama cells, by 0.050 to 0.138 quadratic weighted kappa, five of the six intervals excluding zero. The three Gemma cells at k=8 are also worse, by 0.040 to 0.073, two of the three excluding zero. At $k = 4$ the two Gemma sources are indistinguishable, so Section 4.8 rests its recommendation on absolute quality. The severity profile of Section 4.4 yields a further prediction. Cross-item demonstrations at k=4 display nothing above score 3, so an already lenient model should gain and an already strict one should lose. The sign of the change from zero-shot matches each model's zero-shot bias for all six models at k=4 and for five of six at k=8, the exception being the nationally adapted Gemma model. Rank correlation between bias and change is +0.94, $p = 0.005$, at $k = 4$ and +0.77, $p = 0.072$, at $k = 8$.

The unit is the model, so each value of k rests on six observations. Dropping each model in turn leaves the coefficient between +0.90 and +1.00 at k = 4 and between +0.60 and +0.90 at k = 8. The agreement at four shots does not depend on any single model, while the weaker relation at eight shots does. This supports the severity account of Section 4.7 with data it was not fitted to, not the representational change behind it.

Table 10. Follow-up conditions against their matched cells

Experiment	Family	k	Matched condition	Follow-up condition	Difference [95% CI]
Source question	llama	4	+0.069	+0.005	−0.064
			[+0.042, +0.094]	[−0.032, +0.043]	[−0.107, −0.022]
Source question	gemma	4	−0.047	−0.028	+0.019
			[−0.074, −0.021]	[−0.058, +0.002]	[−0.015, +0.053]
Source question	llama	8	+0.051	−0.014	−0.065
			[+0.004, +0.095]	[−0.071, +0.047]	[−0.140, +0.008]
Source question	gemma	8	−0.018	−0.052	−0.034
			[−0.041, +0.004]	[−0.085, −0.018]	[−0.070, +0.004]
Score coverage	llama	4	+0.069	+0.029	−0.040
			[+0.042, +0.094]	[+0.004, +0.055]	[−0.068, −0.010]
Score coverage	gemma	4	−0.047	−0.009	+0.037
			[−0.074, −0.021]	[−0.050, +0.039]	[−0.007, +0.088]

4.11. Limitations

The findings concern short answer grading, not essay scoring: responses average 10.1 words in the science domain and 12.2 in the social science domain, and no claim is made about longer texts. The evidence rests on 36 independent questions in two subject domains, six of them held out, with subject domain nested completely within question. Claims about generalization to new questions are bounded by those 36, not by the 30,960 grading calls they carry, and claims about subject domain are bounded by two, so the absence of a rung by domain interaction cannot be distinguished from insufficient power. Any statement beyond Biology and Geography examination items is exploratory.

Approximately one per cent of evaluated responses have a text-identical counterpart in the shot pool carrying a different score, so under few-shot a prompt can contain a demonstration that contradicts the reference label of the response being graded. This is the label noise of Section 3.2 and is not corrected for. All generative models were served with 4-bit rather than full-precision weights, identically across rungs, so the comparison remains internally fair but the pattern is not established at full precision. Quantized weights for the unadapted and nationally adapted models came from community repositories, whereas the regional models were published by their own developers, an asymmetry noted although quantization is deterministic given the source weights. Low-rank fine-tuning of the 8- and 9-billion-parameter models was not performed, as it requires full-precision weights and hardware beyond the free-tier accelerator this protocol was designed for. That is the most substantial gap in the design, and L3 provides one fine-tuned data point that is not size-matched.

5. Conclusions

Language adaptation for Indonesian short answer grading does not accumulate with the number of stages applied, and where it stops paying depends on the base model and on the labelled examples available. Regional adaptation was the safe step in both families, whereas national continued pretraining was worth adding only on the weaker base. Labelled examples absorb part of that benefit, so the value of a further stage shrinks for institutions that already hold graded answers, and the source of those answers matters for grading quality as well as for the size of the advantage.

What adaptation changes is where on the score scale a grader is calibrated rather than how well it discriminates, and the direction follows the base model rather than the adaptation procedure. Its value is therefore bounded by the miscalibration the base model leaves to correct, which makes adaptation useful for separating non-answers from answers and

counterproductive for discriminating among adequate ones on the weaker base. The representational change behind this recalibration is not established by the present data.

In practice, generative graders are most valuable when labelled answers are scarce, whereas a fine-tuned Indonesian encoder is the better choice where labels already exist and hardware is constrained. Whether adaptation improves how readily a model can be taught the task, as distinct from how it performs when prompted, remains open and requires full-precision low-rank fine-tuning across the ladder.

Author Contributions: Conceptualization: I.R.B., H.J., and N.; Methodology: I.R.B.; Software: I.R.B.; Validation: I.R.B., H.J., and N.; Formal analysis: I.R.B.; Investigation: I.R.B.; Resources: I.R.B.; Data curation: I.R.B.; Writing—original draft preparation: I.R.B.; Writing—review and editing: H.J. and N.; Visualization: I.R.B.; Supervision: H.J. and N.; Project administration: I.R.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kemdiktisaintek) through Research Grant Number: 231.32/DST/UN34.9/T/PT.01.03/2026**.**

Data Availability Statement: The corpus of Haidir and Purwarianti is publicly available at https://github.com/AgeMagi/tugas-akhir. The repository carries no tagged release, and the contents of its data directory were replaced on 10 October 2026, so every figure reported here reads commit 90ed4c13a5e5b30c29bc209e83dfc59ea0b60e1a, the repository's first commit and the state the experiments used. Work reproducing this article should request that commit rather than the default branch. The corpus is also catalogued as SEACrowd/id_short_answer_grading [38], whose loader resolves to the same source files and whose licence SEACrowd records as unknown. This study reads the source files directly so that the cleaning and per-split merge of Section 3.2 remain visible. Per-unit outputs were retained for every experimental condition and for the fine-tuned encoder, together with the aggregated summary, the contrast tables at both levels of inference, the demonstration profile of each selection rule, the robustness table and a run manifest.

Acknowledgments: The authors thank the developers of the corpus and of the regional and Indonesian-adapted model families for making their resources publicly available.

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

Appendix A. Consolidated configuration

Every setting needed to reproduce the reported runs is gathered here. Checkpoints are identified by repository, the Q4_K_M quantization, and the first 24 hexadecimal characters of the file's SHA-256 digest. The runtime version was not pinned at installation, and the session was not preserved, so it is reconstructed rather than recorded.

A.1 Checkpoints, Quantization, runtime, decoding, and seeds

Setting	Value
llama/L0	QuantFactory/Meta-Llama-3-8B-Instruct-GGUF, 50d2777b87da466eb5bce767
llama/L1	aisingapore/Llama-SEA-LION-v2-8B-IT-GGUF, caea57de8184cfdebf0962e6
llama/L2	gmonsoon/llama3-8b-cpt-sahabatai-v1-instruct-GGUF, feefaa489d7f8eca165fea7a
gemma/L0	bartowski/gemma-2-9b-it-GGUF, f597d13cf2cbc48adc22e2d5
gemma/L1	aisingapore/Gemma-SEA-LION-v3-9B-IT-GGUF, 863b0f186bb363d5848e3145
gemma/L2	gmonsoon/gemma2-9b-cpt-sahabatai-v1-instruct-GGUF, 018e4ee3f2183c58dbc63052
L3	indobenchmark/indobert-base-p1, full precision, fine-tuned, c2cd0b51ddce
Quantization	Q4_K_M, identical for all six generative checkpoints
Runtime	llama.cpp through llama-cpp-python 0.3.34, CUDA 12.4 wheel index, all layers offloaded
Decoding	context 4096 tokens, temperature 0.0, top-p 1.0, maximum 6 output tokens, seed 42

Setting	Value
Sampling seed	42, for both the evaluation set and the shot pool
System role	not used; the rubric is prepended to the first user turn for every model
Hardware	NVIDIA Tesla T4, 16 GB, Google Colab
Encoder L3	3 epochs, batch 16, learning rate 2e-5, maximum length 256, fp16
Baseline B1	TF-IDF, character n-grams within word boundaries, n from 2 to 5, minimum document frequency 2, sublinear term frequency, fitted on training text only
Baseline B2	IndoBERT mean-pooled embeddings, maximum length 128, no fine-tuning
Baseline regressor	ordinary least squares on one cosine-similarity feature, fitted per item, capped conditions averaged over seeds 0, 1, 2

A.2 Prompt

The rubric is written in Indonesian and held identical across all models and conditions. Each response is presented with the item template below. Under few-shot the rubric is followed by k demonstration pairs drawn from the shot pool of the same item, each an item template block followed by its reference score, before the response being graded.

Anda menilai jawaban singkat siswa terhadap satu pertanyaan. Skala penilaian 0 sampai 5: 5 = jawaban benar dan lengkap, mencakup seluruh isi kunci jawaban. 4 = jawaban benar, hampir lengkap, ada satu unsur kecil yang kurang. 3 = jawaban benar sebagian besar, ada unsur penting yang kurang. 2 = jawaban benar sebagian kecil, sebagian besar unsur kunci tidak ada. 1 = jawaban menyinggung topik tetapi isinya tidak tepat. 0 = jawaban salah, kosong, atau tidak berhubungan dengan pertanyaan. Jawab HANYA dengan satu angka dari 0 sampai 5. Jangan menulis apa pun selain angka itu.

Pertanyaan: {item stem} / Kunci jawaban: {expert reference answer} / Jawaban siswa: {student response}

References

[1] M. H. Haidir and A. Purwarianti, "Short Answer Grading Using Contextual Word Embedding and Linear Regression," *J. Linguist. Komputasional*, vol. 3, no. 2, pp. 54–61, 2020, doi: 10.26418/jlk.v3i2.38.

[2] H. R. Salim, C. De, N. D. Pratamaputra, and D. Suhartono, "Indonesian automatic short answer grading system," *Bull. Electr. Eng. Informatics*, vol. 11, no. 3, pp. 1586–1603, Jun. 2022, doi: 10.11591/eei.v11i3.3531.

[3] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.

[4] W. Tjhi, D. Ong, and P. Limkonchotiwat, "SEA-LION (Southeast Asian Languages In One Network): A Family of Southeast Asian Language Models," in *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, 2023, pp. 245–245. doi: 10.18653/v1/2023.nlposs-1.26.

[5] S. Cahyawijaya *et al.*, "Cendol: Open Instruction-tuned Generative Large Language Models for Indonesian Languages," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 14899–14914. doi: 10.18653/v1/2024.acl-long.796.

[6] A. Grattafiori *et al.*, "The Llama 3 Herd of Models," *Arxiv preprint*. Nov. 23, 2024. [Online]. Available: http://arxiv.org/abs/2407.21783

[7] Gemma Team *et al.*, "Gemma 2: Improving Open Language Models at a Practical Size," *Arxiv preprint*. Oct. 02, 2024. [Online]. Available: http://arxiv.org/abs/2408.00118

[8] W. Li and H. Liu, "Applying large language models for automated essay scoring for non-native Japanese," *Humanit. Soc. Sci. Commun.*, vol. 11, no. 1, p. 723, Jun. 2024, doi: 10.1057/s41599-024-03209-9.

[9] J. Schneider, R. Richner, and M. Riser, "Towards Trustworthy AutoGrading of Short, Multi-lingual, Multi-type Answers," *Int. J. Artif. Intell. Educ.*, vol. 33, no. 1, pp. 88–118, Mar. 2023, doi: 10.1007/s40593-022-00289-z.

[10] U. Padó, Y. Eryilmaz, and L. Kirschner, "Short-Answer Grading for German: Addressing the Challenges," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 4, pp. 1321–1352, Dec. 2024, doi: 10.1007/s40593-023-00383-w.

[11] W. Q. Leong, J. G. Ngui, Y. Susanto, H. Rengarajan, K. Sarveswaran, and W. C. Tjhi, "BHASA: A Holistic Southeast Asian Linguistic and Cultural Evaluation Suite for Large Language Models," *Arxiv preprint*. Sep. 19, 2023. [Online]. Available: http://arxiv.org/abs/2309.06085

[12] F. Koto, N. Aisyah, H. Li, and T. Baldwin, "Large Language Models Only Pass Primary School Exams in Indonesia: A Comprehensive Test on IndoMMLU," *arXiv*. Oct. 21, 2023. [Online]. Available: http://arxiv.org/abs/2310.04928

[13] S. Burrows, I. Gurevych, and B. Stein, "The Eras and Trends of Automatic Short Answer Grading," *Int. J. Artif. Intell. Educ.*, vol. 25, no. 1, pp. 60–117, Mar. 2015, doi: 10.1007/s40593-014-0026-8.

[14] M. Mohler and R. Mihalcea, "Text-to-text semantic similarity for automatic short answer grading," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09*, 2009, pp. 567–575. doi: 10.3115/1609067.1609130.
[15] M. O. Dzikovska, R. D. Nielsen, and C. Leacock, "The joint student response analysis and recognizing textual entailment challenge: making sense of student responses in educational applications," *Lang. Resour. Eval.*, vol. 50, no. 1, pp. 67–93, Mar. 2016, doi: 10.1007/s10579-015-9313-8.
[16] K. Taghipour and H. T. Ng, "A Neural Approach to Automated Essay Scoring," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1882–1891. doi: 10.18653/v1/D16-1193.
[17] L.-H. Chang and F. Ginter, "Automatic Short Answer Grading for Finnish with ChatGPT," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 21, pp. 23173–23181, Mar. 2024, doi: 10.1609/aaai.v38i21.30363.
[18] B. R. P. Darnoto and D. B. Firmawan, "Language-Similarity-Guided Transfer Fine-Tuning of Pre-trained Transformer Models for Sentiment Analysis Across 12 Indonesian Regional Languages," *J. Comput. Theor. Appl.*, vol. 3, no. 4, pp. 547–563, May 2026, doi: 10.62411/jcta.15975.
[19] S. Gururangan *et al.*, "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8342–8360. doi: 10.18653/v1/2020.acl-main.740.
[20] GoToCompany, "Sahabat-AI: llama3-8b-cpt-sahabatai-v1-instruct and gemma2-9b-cpt-sahabatai-v1-instruct." 2024.
[21] G. Kortemeyer, "Performance of the pre-trained large language model GPT-4 on automated short answer grading," *Discov. Artif. Intell.*, vol. 4, no. 1, p. 47, Jul. 2024, doi: 10.1007/s44163-024-00147-y.
[22] Y. Kumar, S. Parekh, S. Singh, J. J. Li, R. R. Shah, and C. Chen, "Automatic Essay Scoring Systems Are Both Overstable And Oversensitive: Explaining Why And Proposing Defenses," *Dialogue & Discourse*, vol. 14, no. 1, pp. 1–33, Apr. 2023, doi: 10.5210/dad.2023.101.
[23] A. Filighera, S. Ochs, T. Steuer, and T. Tregel, "Cheating Automatic Short Answer Grading with the Adversarial Usage of Adjectives and Adverbs," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 2, pp. 616–646, Jun. 2024, doi: 10.1007/s40593-023-00361-2.
[24] L. Rodrigues, C. Xavier, N. Costa, D. Gasevic, and R. F. Mello, "Is GPT-4 fair? An empirical analysis in automatic short answer grading," *Comput. Educ. Artif. Intell.*, vol. 8, no. June, p. 100428, Jun. 2025, doi: 10.1016/j.caeai.2025.100428.
[25] C. Grévisse, "LLM-based automatic short answer grading in undergraduate medical education," *BMC Med. Educ.*, vol. 24, no. 1, p. 1060, Sep. 2024, doi: 10.1186/s12909-024-06026-5.
[26] E. Emirtekin and Y. Özarslan, "Automatic Short-Answer Grading in Sustainability Education: <scp>AI</scp> –Human Agreement," *J. Comput. Assist. Learn.*, vol. 42, no. 1, Feb. 2026, doi: 10.1002/jcal.70160.
[27] H. Misgna, B.-W. On, I. Lee, and G. S. Choi, "A survey on deep learning-based automated essay scoring and feedback generation," *Artif. Intell. Rev.*, vol. 58, no. 2, p. 36, Dec. 2024, doi: 10.1007/s10462-024-11017-5.
[28] R. Weegar and P. Idestam-Almquist, "Reducing Workload in Short Answer Grading Using Machine Learning," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 2, pp. 247–273, Jun. 2024, doi: 10.1007/s40593-022-00322-1.
[29] H. Jati, Y. Indrihapsari, S. Romadhon, P. Setialana, D. Wijaya, and S. A. Ardy, "Closed-Source vs. Open-Weights Language Models: Accuracy, Stability, and Deployment Trade-Offs in Educational Assessment (GPT-4o vs. LLaMA 3.1)," *J. Soft Comput. Data Min.*, vol. 7, no. 2, Jun. 2026, doi: 10.30880/jscdm.2026.07.02.005.
[30] L. Jiang and N. Bosch, "Short answer scoring with GPT-4," in *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, Jul. 2024, pp. 438–442. doi: 10.1145/3657604.3664685.
[31] O. Henkel, L. Hills, A. Boxer, B. Roberts, and Z. Levonian, "Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education," in *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, Jul. 2024, pp. 300–304. doi: 10.1145/3657604.3664693.
[32] L. Cong, L. Hammerla, S. Hahn, S. Gombert, H. Drachsler, and U. Kroehne, "Automatic Short Answer Grading with LLMs: From Memorization to Reasoning," in *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, Apr. 2026, pp. 75–84. doi: 10.1145/3785022.3785031.
[33] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," in *arXiv*, Jul. 2020. [Online]. Available: http://arxiv.org/abs/2005.14165
[34] S. Min *et al.*, "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 11048–11064. doi: 10.18653/v1/2022.emnlp-main.759.
[35] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.," *Psychol. Bull.*, vol. 70, no. 4, pp. 213–220, 1968, doi: 10.1037/h0026256.
[36] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *Ann. Stat.*, vol. 7, no. 1, pp. 1–26, Jan. 1979, doi: 10.1214/aos/1176344552.
[37] A. Colin Cameron and D. L. Miller, "A Practitioner's Guide to Cluster-Robust Inference," *J. Hum. Resour.*, vol. 50, no. 2, pp. 317–372, 2015, doi: 10.3368/jhr.50.2.317.
[38] H. Lovenia *et al.*, "SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 5155–5203. doi: 10.18653/v1/2024.emnlp-main.296.