IJECBE (2025), **3**, **3**, 620–650
Received (16 June 2025) / Revised (7 July 2025)
Accepted (7 July 2025) / Published (30 September 2025)
https://doi.org/10.62146/ijecbe.v3i3.148
https://ijecbe.ui.ac.id
ISSN 3026-5258

International Journal of Electrical, Computer and Biomedical Engineering

RESEARCH ARTICLE

Defying Data Scarcity: High-Performance Indonesian Short Answer Grading via Reasoning-Guided Language Model Fine-Tuning

Muhammad Naufal Faza*, Prima Dewi Purnamasari, and Anak Agung Putri Ratna

Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Depok, Indonesia
*Corresponding author. Email: muhammad.naufal428@ui.ac.id

Abstract

Automated Short Answer Grading (ASAG) is crucial for scalable feedback, but applying it to low-resource languages like Indonesian is challenging. Modern Large Language Models (LLMs) severely overfit small, specialized educational datasets, limiting utility. This study compares nine traditional machine learning models against two fine-tuning strategies for Gemma-3-1b-it on an expanded Indonesian ASAG dataset (n=220): (a) standard fine-tuning predicting only scores, and (b) a proposed reasoning-guided approach where the model first generates a score rationale using knowledge distillation before predicting the score. The reasoning-guided model (Gemma-3-1b-ASAG-ID-Reasoning or G-R) achieved state-of-the-art performance (QWK 0.7791; Spearman's 0.8276), significantly surpassing the best traditional model in this study (SVR, QWK 0.6952). This work advances foundational LSA-based approaches for this task by introducing a more robust methodology and evaluation framework. Crucially, standard fine-tuning (Gemma-3-1b-ASAG-ID or G) suffered catastrophic overfitting (QWK 0.7279), indicated by near-perfect training but poor test scores. While the reasoning-guided LLM showed superior accuracy, it required over 35 times more inference time. Results demonstrate that distilled reasoning acts as a powerful regularizer, compelling the LLM to learn underlying grading logic rather than memorizing pairs, establishing a viable method for high-performance ASAG in data-scarce environments despite computational tradeoffs.

Keywords: Automated Short Answer Grading (ASAG), Large Language Models, Reasoning-Guided Fine-Tuning

1. INTRODUCTION

The integration of Artificial Intelligence (AI) into modern educational practices is rapidly transforming pedagogical approaches and administrative efficiencies. Among the diverse applications of AI in education, Automated Short Answer Grading (ASAG) has emerged as a particularly impactful technology [1]. ASAG systems offer the potential to provide students with scalable, consistent, and timely feedback, thereby alleviating the substantial workload traditionally borne by instructors [2]. The capacity of ASAG to process unlimited submissions without additional human resources makes it invaluable for large student cohorts where manual grading would be logistically prohibitive [3]. Furthermore, automated systems can enhance objectivity by applying identical criteria to every submission, mitigating the variability and potential unconscious biases that can influence human graders [4]. Research has documented significant grading inconsistencies in traditional assessment methods, with studies revealing sequential bias where assignments graded later in sequences receive systematically lower scores [5]. The halo effect, where prior positive or negative impressions of students influence subsequent grading decisions, has been empirically demonstrated to create unfair assessment outcomes [6]. Additionally, grading fatigue has been shown to impair judgment quality over time, with teachers demonstrating decreased assessment accuracy as mental fatigue accumulates during extended grading sessions [7]. The provision of rapid feedback is another cornerstone of ASAG's utility, as immediate assessment outcomes are crucial for effective student learning cycles [8], a stark contrast to traditional grading timelines where delays can diminish pedagogical impact. By automating a significant portion of the grading process, ASAG can free educators from time-consuming tasks, allowing them to redirect their efforts towards more personalized student interactions, curriculum refinement, and other higher-value pedagogical activities. Studies have documented substantial time savings, with computer-assisted grading rubrics demonstrating efficiency improvements of 200-350% compared to traditional manual grading methods [3]. This shift suggests that ASAG is not merely an efficiency tool but can act as a catalyst for pedagogical innovation. The ability to generate detailed analytics on student performance across large cohorts and facilitate adaptive learning pathways indicates that ASAG can empower educators with data-driven insights, fostering more dynamic, responsive, and personalized teaching methodologies [9]. Recent advances in large language models have further enhanced ASAG capabilities, with systems achieving grading accuracy comparable to human evaluators while providing comprehensive feedback [10].

Despite the transformative potential of ASAG, its widespread adoption and optimal performance are hindered by two primary obstacles: the "resource gap" and the "data dilemma." The resource gap refers to the pronounced concentration of Natural Language Processing (NLP) research, development, and available resources on high-resource languages, predominantly English [11]. This leaves languages such as Indonesian, spoken by millions, significantly underserved in terms of advanced NLP tools and datasets [12]. The scarcity of comprehensive datasets for languages like Indonesian poses a formidable barrier to building robust NLP applications [13]. Compounding this issue is the "data dilemma" inherent in leveraging state-of-the-

art models, particularly Large Language Models (LLMs). LLMs represent the cutting edge for many NLP tasks due to their sophisticated understanding and text generation capabilities [14]. However, these models are notoriously "data-hungry." When applied to small datasets, a common scenario in specialized real-world educational applications (e.g., answer sets from a single academic course), LLMs are highly prone to overfitting [15]. Overfitting occurs when a model learns the noise and specific details of the training data rather than generalizable patterns, leading to excellent performance on training data but poor performance on unseen data. LLMs, with their vast number of parameters, "can indeed overfit a small dataset" because their high capacity allows them to "memorize examples" instead of learning the underlying features [16]. These two challenges, linguistic resource scarcity and the data-intensive nature of LLMs, create a compounded problem for developing advanced ASAG systems in contexts like Indonesia. Educational datasets are often niche and small by nature; for a low-resource language, this scarcity is amplified, making the successful deployment of powerful LLMs exceptionally difficult without specialized mitigation strategies.

To address this amplified challenge, this paper develops and evaluates a robust, highperforming ASAG system for the Indonesian language within a pragmatically low-resource context. We investigate whether a novel fine-tuning strategy can overcome the critical issue of model overfitting, which severely limits the utility of standard LLM approaches when applied to small, specialized educational datasets. This work aims to establish a viable methodology for deploying powerful language models in data-scarce environments.

This research's primary contribution is the introduction and validation of a novel fine-tuning methodology that leverages distilled reasoning as an effective structural regularizer for Large Language Models. We demonstrate that this "reasoning as regularization" approach directly addresses the critical challenge of catastrophic overfitting, a common failure point when applying standard LLMs to specialized, low-resource datasets. To validate this method, our reasoning-guided model was benchmarked against ten other approaches—including traditional machine learning baselines and a standard fine-tuned LLM—on an expanded Indonesian ASAG dataset. The results show that our model not only avoids the severe overfitting of its standard counterpart but also achieves state-of-the-art performance in grading accuracy and rank-order consistency. Finally, the study provides a critical analysis of the performance-cost trade-offs, quantifying the substantial computational requirements of the proposed method, which is vital information for real-world implementation.

2. Related Works

This section provides a conceptual background for the technologies and methodologies employed in this study, positioning our work within the broader landscape of ASAG and NLP research.

2.1 Traditional and Embedding-based ASAG

Early approaches to Automated Short Answer Grading often relied on extracting explicit linguistic features or employing corpus-based semantic similarity techniques. Among these, Latent Semantic Analysis (LSA) gained prominence [17]. LSA utilizes

mathematical methods to infer relationships between words and documents by analyzing their contextual usage within large text corpora. It represents both words and text passages as vectors in a high-dimensional semantic space, allowing for the computation of semantic similarity even in the absence of direct keyword overlap. For instance, LSA could identify the semantic connection between phrases like "no more dinner" and "skip breakfast" by recognizing the semantic relatedness of "dinner" and "breakfast" as meals [18]. However, while LSA proved effective for scoring longer essays, its application to short, open-ended answers encountered challenges. These included the limited amount of text available for analysis in short responses and difficulties in accommodating the complexity and nuance often present in such answers. Furthermore, the performance of LSA-based systems is inherently tied to the quality of the human ratings used in their training.

The limitations of early feature-based methods spurred a transition towards the use of dense vector embeddings to capture richer and more nuanced semantic similarities between student answers and reference texts. A significant advancement in this area was the development of Sentence-BERT (SBERT) [19]. SBERT modifies the BERT architecture to derive meaningful, fixed-size sentence embeddings. It typically employs a Siamese or biencoder structure, where two sentences are processed independently and in parallel through identical BERT networks. The resulting token embeddings are then pooled (e.g., using mean pooling) to produce a single vector representation for each sentence. This architecture is considerably more efficient for sentence similarity tasks than traditional BERT cross-encoder setups, which require feeding both sentences simultaneously into the model and exhibit quadratic computational complexity for pairwise comparisons. The SentenceTransformers library [20] provides a widely adopted and practical framework for implementing SBERT and similar models, facilitating tasks such as semantic textual similarity and information retrieval. In the context of the present study, sentence embeddings derived from such architectures form the core features for the traditional machine learning baseline models.

The evolution from LSA to SBERT illustrates an ongoing pursuit of more sophisticated semantic understanding in ASAG. LSA offered an initial step beyond keyword matching, but SBERT, leveraging the power of transformer architectures, provided more contextually rich and potent embeddings for capturing subtle semantic relationships. However, even advanced similarity metrics derived from SBERT primarily focus on semantic overlap. The task of grading, particularly for complex subjects, often requires inferential steps, evaluation of reasoning, and assessment of completeness and accuracy that go beyond mere similarity. This sets the stage for investigating more powerful generative models capable of these deeper levels of analysis.

2.2 Generative Large Language Models in NLP

The field of Natural Language Processing has witnessed a significant paradigm shift from task-specific models, such as BERT, to large-scale, generative models. BERT (Bidirectional Encoder Representations from Transformers) [21] was a revolutionary development, primarily due to its ability to understand context bidirectionally by pre-training on masked language modeling and next sentence prediction tasks. This

made BERT and its variants exceptionally effective as encoders for tasks requiring deep contextual understanding, such as text classification, named entity recognition, and the generation of high-quality word and sentence embeddings.

However, the landscape has increasingly been dominated by generative Large Language Models (LLMs), including OpenAI's Generative Pre-trained Transformer (GPT) [22] series and Google's Gemma 3 family [23]. These models are typically characterized by their decoder-only, autoregressive architectures. They are pre-trained on vast amounts of text data to predict the next token in a sequence, a process that endows them with a remarkable ability to generate coherent, contextually relevant, and human-like text based on input prompts or instructions. This generative capability, coupled with their capacity for zero-shot and few-shot learning (i.e., performing tasks with minimal or no task-specific examples) and sophisticated instruction following , makes them prime candidates for complex and nuanced tasks like ASAG. Such tasks may involve not only assigning a score but also generating explanations, providing detailed feedback, or evaluating the reasoning presented in a student's answer.

The move towards generative LLMs like GPT and Gemma for ASAG is motivated by the understanding that grading often demands more than semantic similarity assessment (a strength of SBERT-like models) or text classification (a strength of BERT-like encoders). Effective grading frequently requires the interpretation of instructions, evaluation against multifaceted criteria, and, ideally, the generation of justifications or feedback, all tasks at which generative models excel. The reasoning-guided approach proposed in this study, for instance, explicitly requires the model to generate an analytical rationale. This necessitates the generative prowess offered by models like Gemma 3, distinguishing them from earlier encoder-focused architectures.

2.3 Parameter-Efficient Fine-Tuning (PEFT)

The immense scale of modern LLMs, often comprising billions or even trillions of parameters, presents a significant practical challenge: fully fine-tuning these models for every new downstream task or dataset is computationally prohibitive for most researchers and organizations. This process demands substantial GPU memory, vast amounts of task-specific data, and extensive training time.

Parameter-Efficient Fine-Tuning (PEFT) [24] encompasses a family of techniques developed to address this challenge. PEFT methods enable the adaptation of large pre-trained models to specific tasks by fine-tuning only a small subset of the model's parameters, or by introducing a small number of new, trainable parameters, while keeping the vast majority of the original model weights frozen. This approach offers several key advantages: significantly reduced GPU memory requirements (often allowing fine-tuning on consumer-grade hardware), faster training iterations, smaller storage footprints for the resulting adapted models (as only the small set of changed parameters needs to be saved), and a reduced risk of "catastrophic forgetting," where the model loses its general capabilities learned during pre-training.

Low-Rank Adaptation (LoRA) is a particularly popular and effective PEFT technique [25]. LoRA operates by injecting trainable, low-rank decomposition matrices

into specific layers of the Transformer architecture (typically the attention layers). Instead of fine-tuning the original weight matrices, LoRA introduces two smaller matrices whose product represents the change in weights. Only these low-rank matrices are trained. Since the rank is chosen to be much smaller than the original dimensions of the original weights, the number of trainable parameters is drastically reduced. The LoRA-specific hyperparameters, rank and alpha (a scaling factor), are specified in this study's methodology, indicating its use.

PEFT techniques, especially LoRA, are not merely efficiency optimizations; they are critical enablers for applying LLMs to specialized, low-resource tasks such as the Indonesian ASAG problem investigated in this paper. Without PEFT, attempting to fine-tune a model like Gemma-3-1b-it on a dataset of only 220 samples would be highly impractical and resourceintensive, likely leading to severe overfitting even if computationally feasible for some. PEFT makes such targeted adaptations accessible and manageable, forming a foundational component of this study's experimental design.

2.4 Eliciting Reasoning in Large Language Models

A growing body of research has demonstrated that the performance of LLMs on complex tasks requiring multi-step deduction can be significantly improved by prompting them to generate intermediate reasoning steps, effectively "thinking step-by-step" before producing a final answer. This approach aims to guide the model through a more structured thought process, mirroring human cognitive strategies for problem-solving.

The seminal work in this area is the introduction of Chain-of-Thought (CoT) prompting by Wei et al. [26]. CoT prompting is a technique that encourages LLMs to articulate a series of intermediate reasoning steps that logically lead to the final answer. For instance, when solving a math word problem, instead of directly outputting the numerical solution, a CoT-prompted LLM would first generate the sequence of calculations and logical inferences required to arrive at that solution. This explicit generation of a "chain of thought" has been shown to markedly improve LLM performance on a variety of tasks, including arithmetic reasoning, commonsense reasoning, and symbolic manipulation. A key benefit of CoT is the increased transparency it offers into the model's decision-making process; the explicit reasoning steps can be inspected to understand how the model reached its conclusion, aiding in debugging and building trust.

The work of Wei et al. [26] on zero-shot and few-shot CoT prompting serves as a direct inspiration for the supervised fine-tuning target employed in the present study. However, this study takes the concept a step further. Instead of relying solely on prompting strategies at inference time to elicit reasoning, our proposed methodology trains the smaller student LLM (Gemma-3-1b-it) to explicitly generate an analytical reasoning step as an integral part of its output. The "ground truth" reasoning for this training is created through knowledge distillation, where a larger, more capable teacher model (DeepSeek R1-0528 [27]) generates these analytical rationales for each sample.

This approach signifies an evolution from CoT as an inference-time prompting

strategy to CoT-like reasoning as an ingrained, learned capability of the model. By making the generation of a coherent rationale a direct supervised learning objective, the aim is to instill a more robust and reliable reasoning faculty within the smaller student model. This contrasts with relying on the emergent reasoning abilities that are typically elicited by specific prompt phrasing in much larger models. The hypothesis is that this supervised learning of reasoning will be more effective for smaller models operating in low-data regimes and, critically, that the process of learning to generate such reasoning will act as a powerful regularizer, which is central to this paper's thesis.

2.5 NLP and ASAG in the Indonesian Context

The Indonesian language (Bahasa Indonesia), despite being the national language of a populous country and a lingua franca for millions, is generally considered a low-resource language in the context of NLP development. While there has been a notable increase in advancements since approximately 2020, a significant disparity remains compared to high resource languages like English. Key models developed for Indonesian include IndoBERT (a BERT-based model for Indonesian understanding tasks) [28], NusaBERT (a multilingual extension of IndoBERT covering Indonesian and several regional languages) [29], and IndoT5 (a T5-based sequence-to-sequence model for Indonesian generation tasks) [30]. Despite this progress, the overall availability of large-scale, high-quality datasets and specialized NLP tools for Indonesian remains limited.

The specific task of Indonesian ASAG, and the dataset it is based on, was foundationally explored by Ratna et al. [31]. Their pioneering research developed an automatic grading system by first using a Support Vector Machine (SVM) to classify answers by topic, then employing Latent Semantic Analysis (LSA) to score the semantic similarity against a reference answer. Evaluated on a 148-sample version of the dataset, their LSA-based scoring system achieved a reported accuracy of 72.01%. This work was crucial in establishing the task's feasibility with classic statistical NLP methods. Our study revisits this regression problem but shifts the paradigm in two significant ways: first, by replacing statistical similarity with generative fine-tuning, and second, by employing a suite of modern evaluation metrics (e.g., QWK, Spearman's) that are specifically designed for assessing ordinal agreement and rank-order consistency in grading tasks.

Other notable contribution include the work by Wijaya [32], who developed an automatic short answer grading system for Indonesian using BERT. Their methodology involved collecting short answers from high school students on Computer and Information Technology subjects, preprocessing the data by concatenating questions with student answers, tokenizing the combined text, and then fine-tuning a BERT model. To manage complexity, they simplified the grading into a binary classification task (correct vs. wrong). Their system achieved a Cohen's Kappa coefficient of 0.75, with high precision (0.94) and recall (0.96) on this binary task, demonstrating the feasibility of using transformer-based models for Indonesian ASAG. Other related work in the Indonesian context includes studies on the utilization of NLP-powered applications like Google Translate and Grammarly by students and educators, and

the availability of general NLP libraries like spaCy with support for Indonesian [33]. The present study builds upon these foundational efforts while aiming to address more profound challenges.

The present study builds directly upon the foundational efforts of Ratna et al. [31] and the broader context provided by Wijaya [32]. However, our research is designed to test a specific hypothesis, which requires a precise evaluation framework and informs our choice of baselines relative to established Indonesian models like IndoBERT, NusaBERT, and IndoT5. Our argument for the superiority of the proposed model is two-fold. First, to justify using a complex LLM, we test whether the entire paradigm is superior to traditional machine learning for this task by benchmarking against a comprehensive suite of nine classical models. This allows us to evaluate whether the LLM approach offers a substantial performance leap over established methods like a highly tuned Support Vector Regressor. Second, and more critically, we aim to prove why our specific fine-tuning method is superior in a low-resource context. The central hypothesis is that using distilled reasoning as a co-training target acts as a powerful regularizer that mitigates the catastrophic overfitting seen in standard LLM fine-tuning. To test this, the most direct and scientifically valid control is an identical base model (Gemma-3-1b-it) fine-tuned without the reasoning component, allowing us to measure the direct effect of our method. This two-part experimental design necessitates our specific choice of baselines; encoder-only models like IndoBERT, while important for Indonesian NLP, are architecturally unsuited for the generative rationale task at the core of our hypothesis and would introduce confounding variables, shifting the focus from a rigorous test of our methodology. Thus, this study advances the state-of-the-art by distinguishing itself in several key aspects:

1. It employs a more recent generative LLM, Gemma-3-1b-it, designed for instruction following and generation.
2. The task is more granular, involving the prediction of a numerical score on a 0-100 scale rather than binary classification.
3. A novel reasoning-guided fine-tuning methodology is introduced, incorporating knowledge distillation from a powerful teacher model to generate analytical rationales.
4. The core investigation focuses on the role of this reasoning generation as a regularization technique to combat the severe overfitting typically encountered when applying LLMs to small datasets, a critical issue in low-resource language contexts.
5. The study utilizes a unique dataset format based on direct professor's answer versus student's answer comparisons.

Previous work, such as that by Wijaya [32], has confirmed the applicability of transformer models for Indonesian ASAG. This study seeks to advance the state of the art by leveraging more sophisticated LLMs for a more complex regression-based grading task. More importantly, it introduces a specific methodological innovation, reasoning-guided fine-tuning, explicitly designed to overcome the fundamental limitation of data scarcity and the resultant LLM overfitting. By demonstrating that this

approach can act as an effective regularizer, this research aims to provide a pathway for developing more accurate and nuanced ASAG systems for Indonesian and potentially other low-resource languages.

3. Methodology

This section details the experimental setup, including the dataset, model implementations, fine-tuning procedures, and evaluation protocol, to ensure the reproducibility of our findings. The overall architecture of our proposed reasoning-guided methodology, encompassing both the knowledge distillation and fine-tuning in the training phase and the final grading in the inference phase, is illustrated in Figure 1.

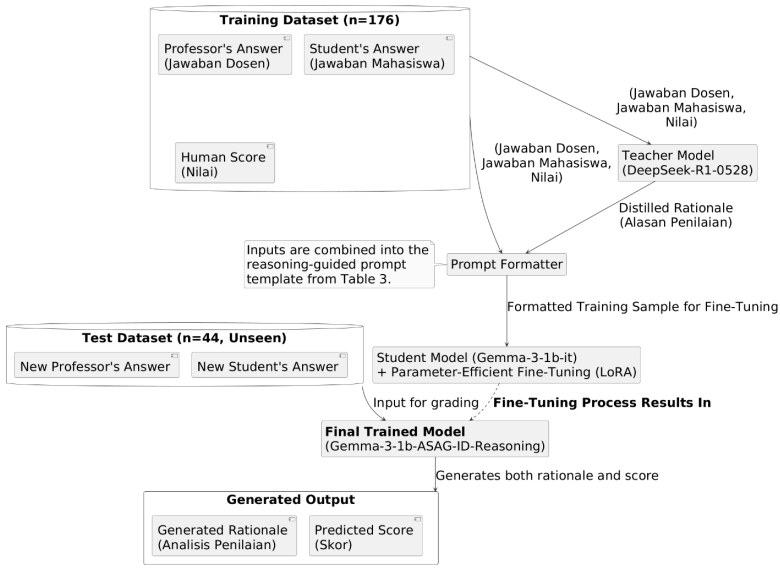

Figure 1. Architecture of the Proposed Reasoning-Guided Fine-Tuning Methodology

3.1 Dataset and Task Formulation

The dataset utilized in this study is an expanded and curated version of the corpus introduced by Ratna et al. [31], now consisting of 220 Indonesian short answer pairs. Each entry comprises a professor's reference answer (jawaban_dosen), a student's submitted answer (jawaban_mahasiswa), and a human-assigned numerical score (nilai) on a scale of 0 to 100. The numerical scores were assigned by a single domain expert to ensure grading consistency across all samples. The descriptive statistics for the scores across the entire dataset are: mean (π) = 75.36, standard deviation (σ) = 24.31, minimum = 0, first quartile (25%) = 60, median (50%) = 80, third quartile (75%) = 100, and maximum = 100. The question-answer pairs in the dataset, while focused, cover several core topics in computer science and engineering, providing thematic diversity. These topics include: 1) Computer Architecture, with questions focused on the foundational Von Neumann architecture and its main components (CPU, ALU,

memory, I/O); 2) Processor Performance Technology, repeatedly explaining Intel's Turbo Boost feature and its function of automatically increasing a processor's clock frequency; 3) Network Models, addressing the client-server model and the respective roles of the client and server; and 4) Operating System Features, with answers detailing the PC hibernation function and how it differs from sleep mode by saving the system's state to non-volatile storage.

The dataset was partitioned into training and testing sets using an 80/20 split, resulting in 176 samples for training and 44 samples for testing. To ensure that both sets were representative of the overall score distribution, a stratified splitting strategy was employed based on score quantiles. The distribution of data across these quantiles and splits is presented in Table 1. This stratification is particularly important for small datasets, as it helps to prevent sampling bias and ensures that the model is trained and evaluated on comparable data distributions, leading to more reliable and generalizable results within the context of this dataset. The task is formulated as a regression problem: given the professor's answer and the student's answer, the models are trained to predict the numerical score.

Table 1. Dataset Score Distribution and Stratified Split

Statistic/Bin	Combined (n=220)	Train (n=176)	Test (n=44)
Mean Score	75.36	74.99	76.82
Std. Dev.	24.31	25.12	21.00
Quantile Bins			
Bin 0 (Scores $\leq$ 60)	64 (29.1%)	51 (29.0%)	13 (29.5%)
Bin 1 (70 $\leq$ Scores $\leq$ 80)	62 (28.2%)	49 (27.8%)	13 (29.5%)
Bin 2 (Scores = 90)	37 (16.8%)	30 (17.0%)	7 (15.9%)
Bin 3 (Scores = 100)	57 (25.9%)	46 (26.1%)	11 (25.0%)

The careful construction and splitting of this dataset are foundational to the study. The stratified split ensures that the evaluation of models, particularly in a low-data regime, is as fair and robust as possible by maintaining proportional representation of score categories across the training and testing subsets. This minimizes the risk that observed performance differences are mere artifacts of skewed data distributions in the splits.

3.2 Baseline Model Implementation

Nine baseline models were implemented to provide a comprehensive comparison against the LLM-based approaches. These models include:

1. **Linear Regression (LinearR):** A standard implementation with no hyperparameter tuning.
2. **Ridge Regression (RR):** Tuned across 5 alpha values.
3. **Lasso Regression (LassoR):** Tuned across 5 alpha values.
4. **Elastic Net Regression (ENR):** Tuned using 3 alpha values and 3 L1_ratios (9 total combinations).

5. Support Vector Regressor (SVR): Tuned across three kernels for a total of 20 combinations:
    a) **Linear Kernel:** 4 C values.
    b) **Polynomial Kernel:** 8 combinations (2 C values, 2 gamma values, 2 degree values).
    c) **RBF Kernel:** 8 combinations (4 C values, 2 gamma values).
6. **Gradient Boosting Regressor (GBR):** Tuned with 8 combinations (2 n_estimators, 2 learning_rates, 2 max_depths).
7. **Random Forest Regressor (RFR):** Tuned with 12 combinations (2 n_estimators, 3 max_depths, 2 min_samples_splits).
8. **K-Neighbors Regressor (KNR):** Tuned across 5 n_neighbors values.
9. **Feed-forward Neural Network (NN):** A specific architecture with layer sizes of [385, 192, 96, 48, 1], 50% dropout in the first three layers, and trained for 40 epochs using an AdamW optimizer.

Feature engineering for these baseline models was conducted as follows: Sentence embeddings for both the student's answer and the professor's answer were generated using the intfloat/multilingual-e5-large-instruct model, a powerful multilingual sentence encoder known for its strong performance on semantic representation tasks. The final feature vector for each answer pair consisted of three components:

- A 1024-dimensional difference vector, calculated as (student_embedding - professor_embedding). This vector aims to capture the semantic divergence between the student's and professor's answers.
- A scalar feature representing the word count of the student's answer.
- A scalar feature representing the word count of the professor's answer.

Hyperparameters for seven of the nine baseline models (Ridge, Lasso, Elastic Net, Support Vector Machine, Gradient Boosting, Random Forest, and K-Neighbors) were optimized using a 5-fold cross-validation procedure on the training set to maximize their performance. The selection of a potent sentence embedding model and the application of hyperparameter tuning for most baselines ensure that these traditional models serve as strong benchmarks. This robust baseline setup allows for a fairer and more credible assessment of any performance gains achieved by the more complex LLM-based approaches.

3.3 LLM-Based Approach

Two distinct LLM-based approaches were developed and evaluated, both centered on the google/gemma-3-1b-it model.

google/gemma-3-1b-it was selected as the base model for fine-tuning due to its favorable performance-to-size ratio among contemporary open-source LLMs. The "-it" suffix denotes an instruction-tuned variant, making it well-suited for tasks that require understanding and responding to specific directives, such as those in ASAG. For the reasoning-guided approach, DeepSeek-R1-0528 was employed as the teacher

model. This model was chosen for its state-ofthe-art open-source LLM at the time of experimentation, making it suitable for generating the high-quality analytical rationales (alasan_penilaian) used as training targets.

Both LLM experiments utilized Parameter-Efficient Fine-Tuning (PEFT) with Quantized Low-Rank Adaptation (LoRA) to make the fine-tuning process computationally tractable and efficient. The key hyperparameter settings, common to both LLM fine-tuning experiments, were:

- LoRA rank: 64
- LoRA alpha: 64 (a scaling factor for LoRA)
- Learning rate: 0.0002
- Maximum sequence length: 1024 tokens
- Quantization: 4-bit quantization was applied to further reduce memory requirements.
- Training epochs: 100
- Batch size: 4
- Gradient accumulation steps: 8 (resulting in an effective batch size of 32)

The prompt structures for both fine-tuning approaches, detailed in Tables 2 and 3, were deliberately engineered to ensure clarity, controllability, and effective guidance for the model. The design incorporates several key principles. First, the prompt initiates with a role-playing instruction ("Anda asisten penilai ahli"), a technique used to prime the model into a specific context, leading to more focused and relevant outputs. Second, it explicitly lists the evaluation criteria (e.g., conformity, completeness, clarity) to direct the model's attention to the specific dimensions of a high-quality answer. Third, the use of XML-like tags (e.g.,<skor>, </skor> , <analisis _ penilaian>) is critical for enforcing a structured output. This not only makes the final score and rationale programmatically easy to parse during inference but is also vital during training for masking the loss calculation to only the target text. This structured approach minimizes ambiguity and ensures the model's responses are consistent and directly usable.

3.3.1 Standard Fine-Tuning: Gemma-3-1b-ASAG-ID (G)

This model serves as a control to assess the performance of the Gemma-3-1b-it model when fine-tuned directly for the scoring task without the reasoning-guidance component. Training samples were formatted using an instructional prompt designed to elicit only the numerical score, as shown in Table 2. The loss calculation was masked to ignore the instructional context, focusing solely on the predicted score. The model's target output (the assistant's message) was structured to contain only the human-assigned score.

3.3.2 Reasoning-Guided Fine-Tuning (Proposed Method): Gemma-3-1b-ASAG-ID-Reasoning (G-R)

This model represents the proposed methodology, where the LLM is trained to generate an analytical rationale before providing the score.

To formalize this process, let us define the inputs as the professor's answer, A_p, and the student's answer, A_s . A standard fine-tuning approach, as described in Section 3.3.1, aims to learn a direct mapping function, f, such that $f(A_p, A_s) \rightarrow S_{predicted}$, where $S_{predicted}$ is the predicted score.

In contrast, our proposed reasoning-guided method introduces an intermediate analytical step. The process begins with knowledge distillation, where a powerful teacher model, $M_{teacher}$, generates a rationale, $R_{teacher}$, based on the full input triplet from the training data, including the human-assigned score, S_{human}, such that $M_{teacher}, (A_p,A_s,S_{human}) \rightarrow R_{teacher}$.

This generated rationale, $R_{teacher}$, then serves as a supervised learning target alongside the human score. The student model is therefore trained to learn a more complex function,g, that produces a tuple containing both a predicted rationale, $R_{predicted}$, and a predicted score, $S_{predicted}$, such that $g(A_p, A_s) \rightarrow (R_{predicted}, S_{predicted})$.

By requiring the model to learn this joint distribution, the rationale generation task $R_{predicted}$ acts as a structural regularizer on the score prediction task ($S_{predicted}$), compelling the model to learn the underlying grading logic rather than simply memorizing input-score pairs.

The DeepSeek-R1-0528 teacher model was first used to generate an analytical reason (*alasan_penilaian*) explaining the human-assigned score for each sample in the training dataset. This *alasan_penilaian* was then incorporated into the training target for the G-R model. The quality of this distilled reasoning is pivotal, as the success of the reasoning-as-regularization approach is fundamentally dependent on the student model being trained on high-fidelity analytical targets. To formally validate the output of the teacher model, a rigorous verification protocol was therefore established and applied to the entire set of 176 generated rationales (*alasan_penilaian*). Each rationale was systematically evaluated against a multi-dimensional quality rubric to ensure it met the highest standards. The criteria included: 1) Logical Congruence, confirming that the justification's analytical depth and sentiment were proportionally aligned with the human-assigned score; 2) Content Specificity, verifying the reasoning was explicitly grounded in the unique substance of the student's answer relative to the reference answer, not on generic templates; and 3) Factual Integrity, ensuring all statements within the rationale were accurate to the subject matter. This comprehensive manual evaluation confirmed that the teacher model consistently produced coherent, generalizable reasoning that reflected the underlying grading logic for every sample in the training set, affirming that the distilled rationales provided a robust and reliable signal for the regularization task.

The model was trained using an instructional prompt that explicitly asked for both an analysis and a score, as shown in Table 3. The model's target output (the assistant's message) was structured to include both the distilled analytical reason and the human-assigned score.

3.4 Evaluation Protocol

Model performance was assessed using a suite of five standard evaluation metrics:

- **Quadratic Weighted Kappa (QWK):** Measures inter-rater agreement for or-

Table 2. Template Structure of the Standard Fine-Tuning

Prompt Structure	**Target Structure**
Anda asisten penilai ahli. Tugas: evaluasi "Jawaban Mahasiswa" vs "Kunci Jawaban Dosen" dengan pertimbangan kesesuaian dengan kunci jawaban, kelengkapan argumen, panjang jawaban, kejelasan bahasa, dan akurasi informasi. Langsung berikan skor akhir dari 0-100 di antara <skor>dan</skor>. Jangan berikan analisis atau teks tambahan apapun Kunci Jawaban Dosen: <kunci _ jawaban> {jawaban_dosen} </kunci_jawaban> Jawaban Mahasiswa: <jawaban_mahasiswa> {jawaban_mahasiswa} Analisis dan Penilaian:	<skor> {human_score} </skor>

dinal scales. It assigns greater penalties to disagreements that are further apart, making it particularly suitable for evaluating grading tasks where the magnitude of score difference matters.

- **Root Mean Squared Error (RMSE):** Calculates the square root of the average of the squared differences between predicted and actual scores. RMSE is sensitive to large errors.
- **Mean Absolute Error (MAE):** Computes the average of the absolute differences between predicted and actual scores. MAE is less sensitive to outliers than RMSE.
- **Pearson's Correlation Coefficient (r):** Measures the linear correlation between predicted and actual scores.
- **Spearman's Rank Correlation Coefficient ():** Measures the monotonic relationship between predicted and actual scores by comparing their ranks. It is less sensitive to the specific distribution of scores than Pearson's correlation and is valuable for assessing if models rank answers in a similar order to human graders.

The use of this comprehensive set of metrics allows for a multifaceted understanding of model performance. While RMSE and MAE quantify prediction error, QWK and Spearman's correlation provide insights into the agreement with human judgment on an ordinal scale and rank-order consistency, respectively. These aspects are often more reflective of true grading quality than raw error metrics alone, as a model might achieve low MAE but still exhibit poor QWK if its errors, though small on average, consistently cross important grade boundaries or fail to preserve the relative ranking of answers.

Table 3. Template Structure of the Reasoning-Guided Fine-Tuning

Prompt Structure	Target Structure
Anda asisten penilai ahli. Tugas: evaluasi "Jawaban Mahasiswa" vs "Kunci Jawaban Dosen" dengan pertimbangan kesesuaian dengan kunci jawaban, kelengkapan argumen, panjang jawaban, kejelasan bahasa, dan akurasi informasi. Beri analisis singkat kelebihan/kekurangan di antara <analisis_penilaian>dan </analisis_penilaian>, lalu akhiri dengan skor 0-100 di antara <skor>dan </skor>. Kunci Jawaban Dosen: <kunci_jawaban> {jawaban_dosen} </kunci_jawaban> Jawaban Mahasiswa: <jawaban_mahasiswa> {jawaban_mahasiswa} </jawaban_mahasiswa> Analisis dan Penilaian:	<analisis_penilaian> {alasan_penilaian} <analisis_penilaian> <skor> {human_score} </skor>

3.5 Implementation Details

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU, an AMD Ryzen 7 5700X CPU, and 32 GB of RAM. The operating system was Arch Linux (x86_64). The primary software libraries and frameworks utilized included Unsloth (for efficient QLoRA fine-tuning of LLMs), Sentence Transformers (for generating sentence embeddings), TRL (Transformers Reinforcement Learning library, used for its supervised fine-tuning utilities), and scikit-learn (for implementing baseline machine learning models and evaluation metrics). Documenting these details is crucial for ensuring the reproducibility of the experiments and providing context for the reported performance benchmarks, such as training and inference times.

4. Results and Analysis

This section presents the empirical findings of the study, beginning with an overall comparison of model performance and computational costs, followed by a study focusing on the regularizing effect of reasoning-guided fine-tuning, and concluding with a qualitative analysis of model behaviors.

4.1 Overall Performance and Computational Cost

Table 4 presents the comprehensive performance metrics and computational costs for all 11 models evaluated on the Indonesian ASAG test set. Following hyper parameter tuning, the best performing configurations for the classical models were used to generate these results. The final parameters were: Ridge (alpha: 0.1), Lasso (alpha: 0.01),

Elastic Net (alpha: 0.1, l1_ratio: 0.1), SVR (C: 50, kernel: 'linear'), Gradient Boosting (learning_rate: 0.1, max_depth: 3, n_estimators: 100), Random Forest (max_depth: 20, min_samples_split: 5, n_estimators: 100), and K-Neighbors (n_neighbors: 5). This table provides the primary quantitative basis for comparing the efficacy of these traditional machine learning approaches against our LLM fine-tuning strategies.

Table 4. Overall Model Performance and Computational Costs on the Test Set (n=44)

Model	QWK	RMSE	MAE	Pearsons	Spearmans	Training Time (s)	Inference Time (s)	RAM Usage (KB)	VRAM Usage (MB)
LinearR	0.5322	16.9727	13.8882	0.5975	0.5425	0.0218	0.0002	9.80	**0.00**
RR	0.6547	13.7230	11.2274	0.7652	0.6847	0.0051	0.0002	**8.45**	**0.00**
LassoR	0.6308	15.3129	12.4269	0.6753	0.6557	0.0339	0.0002	8.53	**0.00**
ENR	0.5216	17.1383	13.8482	0.5697	0.6081	0.0063	**0.0001**	8.53	**0.00**
SVR	0.6952	13.5348	10.8923	**0.7822**	0.7508	0.2184	0.0011	1413.79	**0.00**
GBR	0.6372	14.9279	11.5406	0.6987	0.6795	4.3462	0.0004	131.58	**0.00**
RFR	0.6757	14.8146	11.7906	0.7175	0.7233	5.2229	0.0128	417.24	**0.00**
KNR	0.6621	15.6578	12.4545	0.6629	0.6251	**0.0015**	0.0319	1412.71	**0.00**
NN	0.6024	15.0950	12.3353	0.7034	0.6342	0.3846	0.0003	1924.77	**0.00**
G	0.7279	17.0561	10.9091	0.7040	0.6488	1426.0	1.5480	2857984.00	2989.00
G-R	**0.7791**	**13.1426**	**9.0909**	0.7802	**0.8276**	1510.0	54.6260	2999296.00	10464.00

Analyzing the performance metrics, the G-R model clearly emerges as the state-of-the-art performer on this dataset. It achieved the highest Quadratic Weighted Kappa (QWK) of 0.7791 and the highest Spearman's rank correlation of 0.8276. These metrics, which emphasize agreement on an ordinal scale and rank-order consistency respectively, are particularly crucial for ASAG tasks. This model outperformed the next-best traditional machine learning model, SVR (QWK 0.6952, Spearman 0.7508), surpassed the standard fine-tuned LLM, G (QWK 0.7279, Spearman 0.6488). This result represents a substantial methodological leap over the foundational LSA-based approaches previously applied to this task [31]. By moving from statistical similarity to generative reasoning, our approach is designed to capture more complex nuances of language, a capability reflected in the high QWK and Spearman's scores. These metrics, which are standard for modern grading tasks, provide a more robust assessment of performance than the custom accuracy metrics used in earlier work. The G-R model also achieved the lowest MAE (9.0909) and RMSE (13.1426) among all models, indicating superior accuracy in predicting the absolute scores.

Notably, this superior performance comes at a steep price in terms of computational resources. The G-R model's inference time of 54.6260 seconds for the 44-sample test set is over 35 times longer than that of its non-reasoning counterpart, G (1.5480 seconds). Furthermore, its inference time is several orders of magnitude greater than that of the highly efficient SVR model (0.0011 seconds). The VRAM usage during inference for the reasoning model (10464.00 MB) is also substantially higher than for the non-reasoning LLM (2989.00 MB) and vastly greater than the negligible VRAM usage of traditional models. This highlights a critical performance-versus-efficiency trade-off that practitioners must consider. While the reasoningguided approach yields the highest accuracy, its deployment in real-time or resourceconstrained scenarios would require careful consideration of these computational overheads. The results starkly illustrate that achieving high accuracy in this

challenging low-resource Indonesian ASAG task with LLMs requires not only a sophisticated method like reasoning-guidance but also incurs substantial computational demands.

4.2 The Regularizing Effect of Reasoning-Guided Fine-Tuning

To investigate the impact of the reasoning-guided fine-tuning approach on model generalization, a study was conducted by comparing the training and testing performance of GR (proposed method) against G (standard fine-tuning) over 100 epochs. Figure 2, Figure 3, Figure 4, Figure 5, and Figure 6 visualize the learning curves for RMSE, MAE, Pearson's correlation, Spearman's correlation, and QWK for both models on both training and test sets.

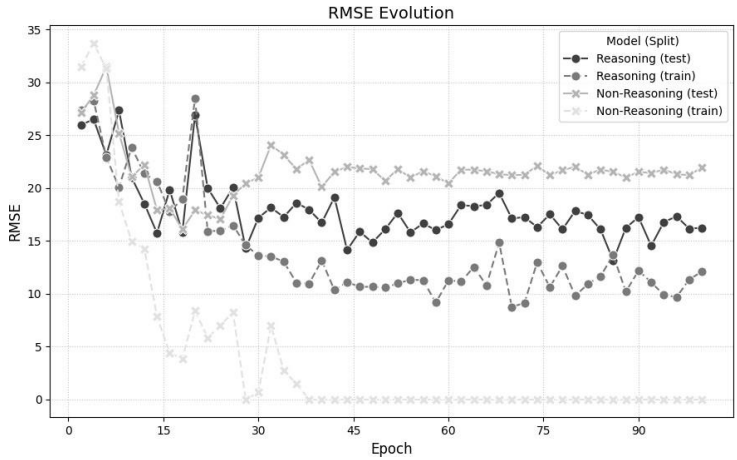

Figure 2. RMSE Evolution of Gemma-3 Fine-Tuning

The behavior of the G model, as depicted by its learning curves, is a textbook example of catastrophic overfitting on a small dataset. The training loss, represented by RMSE and MAE, plummets to near-zero values by epoch 30 (Figure 2, Figure 3), while the training QWK and correlation scores simultaneously saturate at a perfect 1.0 (Figure 4, Figure 5, Figure 6). This indicates that the model has not learned a generalizable grading function but has instead perfectly memorized the 176 training examples. The test performance curves confirm this failure. After a brief period of improvement, the test performance peaks prematurely around epoch 24 and subsequently stagnates or even degrades. This creates a massive and growing divergence between the perfect training scores and the poor test scores, a classic sign that the model is not viable for real-world use on unseen data.

In stark contrast, the G-R model's learning curves demonstrate the powerful regularizing effect of the reasoning-guided approach. Its training error decreases more gradually and stabilizes at a non-zero plateau (e.g., RMSE around 10-12), which is indicative of a model that is learning patterns rather than memorizing noise. Most importantly, its test performance curves, while exhibiting noticeable volatility, track the trend of the training curves far more closely across all metrics. This jaggedness

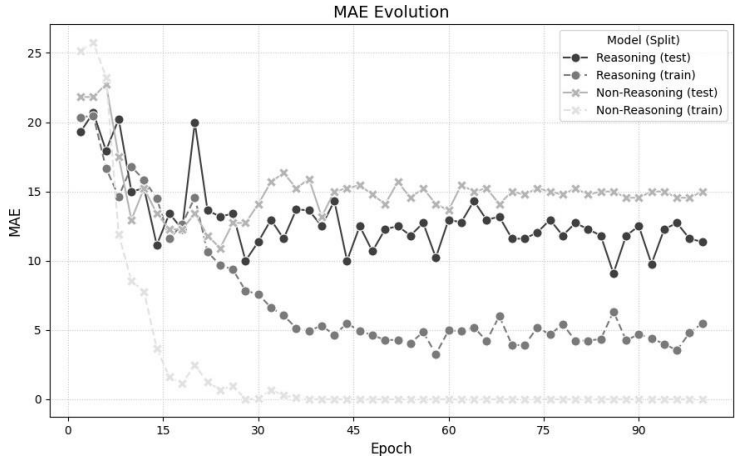

Figure 3. MAE Evolution of Gemma-3 Fine-Tuning

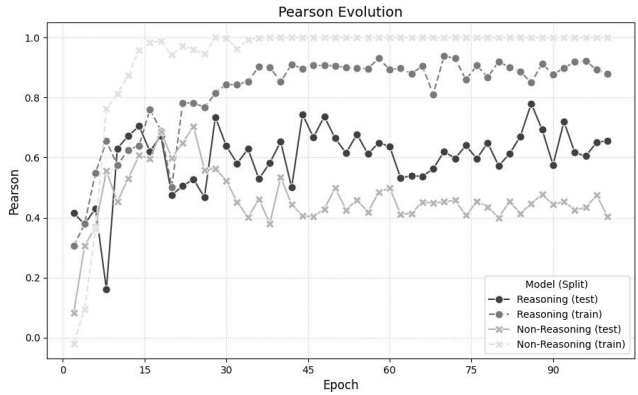

Figure 4. Pearson Evolution of Gemma-3 Fine-Tuning

in the test curves is expected given the small test set size (n=44), where the misclassification of even a few samples can cause significant epoch-toepoch fluctuations in aggregate scores. However, unlike the G model, the G-R model's test performance sustains a stable, upward trajectory long into the training process, eventually plateauing at a much higher level of performance without the catastrophic divergence seen in the standard fine-tuning approach. This results in a significantly smaller generalization gap (Table 5) and demonstrates that the model has learned a more robust and useful grading logic. The requirement to generate a coherent rationale has successfully constrained the model, preventing it from collapsing into a simple memorization strategy and forcing it to develop a more nuanced understanding of the grading task.

To provide a precise numerical summary of the trends observed in the learning curves, Table 5 presents the key performance characteristics of each model. The table highlights the peak performance achieved on the unseen test set and contrasts it

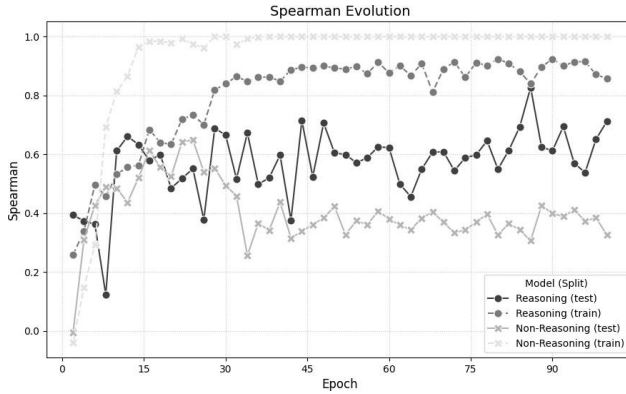

Figure 5. Spearman Evolution of Gemma-3 Fine-Tuning

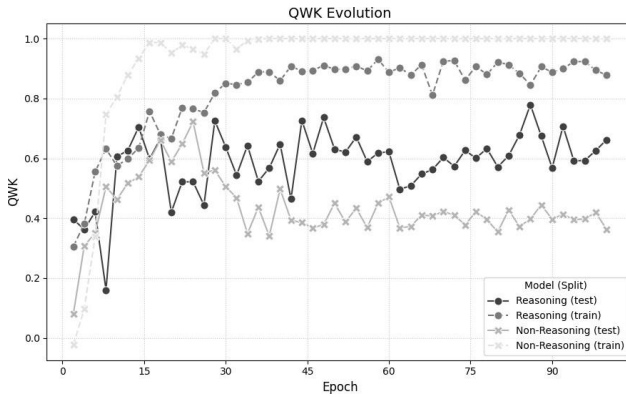

Figure 6. QWK Evolution of Gemma-3 Fine-Tuning

with the training set performance at that same epoch, thereby quantifying the critical generalization gap.

The data in Table 5 crystallizes the findings from the learning curves, offering definitive proof of the non-reasoning model's overfitting. It reaches its peak test performance remarkably early, around epoch 24 for most metrics. At this early stage, a massive generalization gap is already evident. For instance, while its peak test QWK is 0.7222, its training QWK has already soared to 0.9643, creating a substantial gap of 0.2420. This pattern, where the model performs far better on data it has memorized than on new data, confirms that it has prioritized rote learning over developing a true, generalizable grading logic. In stark contrast, the reasoning-guided model demonstrates the hallmarks of effective generalization. It achieves a significantly higher peak test score (e.g., a QWK of 0.7791 and a Spearman correlation of 0.8276) much later in the training, around epoch 86. More importantly, its generalization gap is an order of magnitude smaller across all key metrics.

Table 5. Summary of Model Performance and Generalization Gap at Peak Test Performance

Metric	Statistic	Non Reasoning	Reasoning
QWK	Test Score (Peak)	0.7222	**0.7791**
	Train Score (at Peak)	**0.9643**	0.8448
	Generalization Gap	0.2420	**0.0657**
	Peak Epoch	**24**	86
Spearman	Test Score (Peak)	0.6488	**0.8276**
	Train Score (at Peak)	**0.9740**	0.8405
	Generalization Gap	0.3252	**0.0129**
	Peak Epoch	**24**	86
RMSE	Test Score (Peak)	16.0963	**13.1426**
	Train Score (at Peak)	**3.8552**	13.6614
	Generalization Gap	**12.2411**	**0.5188**
	Peak Epoch	**18**	86
MAE	Test Score (Peak)	10.9091	**9.0909**
	Train Score (at Peak)	**0.6857**	6.3371
	Generalization Gap	10.2234	**2.7538**
	Peak Epoch	**24**	86

The QWK gap is only 0.0657 (compared to 0.2420 for the non-reasoning model), and the Spearman gap is a mere 0.0129. This indicates that the training and test performance curves track each other much more closely, a classic sign of a well-regularized model that learns the task's underlying principles.

These results provide strong evidence that the reasoning-generation step acts as a powerful structural regularizer. By forcing the model to generate a coherent justification for its score, a task that is inherently more complex than merely predicting a numerical value, we constrain the optimization process. This additional constraint prevents the model from collapsing into a simple memorization of (answer, score) pairs from the limited training data. Instead, it compels the model to learn the underlying grading logic embedded within the (distilled) reasoning traces. The requirement to produce a structured textual explanation alongside the score effectively increases the complexity of the learning task. Rather than mapping input text to a single number, the model must now learn a more intricate mapping to both a textual rationale and a score, where the rationale must be consistent with the score and reflect the teacher model's grading logic. This increased task demand makes it more difficult for the model to find trivial solutions that only memorize the scoring subtask, thereby promoting the learning of more generalizable features pertinent to the actual grading criteria.

4.3 In-Depth Diagnostic Analysis of Model Behavior

To move beyond aggregate metrics and gain a deeper understanding of model quality, a series of diagnostic tests were performed. This section presents a quantitative analysis of model prediction errors and an investigation into potential model bias stemming from superficial heuristics like answer length.

4.3.1 Quantitative Error Analysis

The residual plots (predicted score vs. residuals) serve as a primary tool for assessing the validity of a regression model's assumptions. The plot for the SVR model (Figure

7) reveals significant heteroscedasticity. The variance of the residuals is not constant, forming a funnel shape where predictive error is substantially larger for lower predicted scores than for higher ones. This indicates that the model's reliability is inconsistent across its predictive range. The plot for the G model (Figure 8) exhibits an even more severe form of heteroscedasticity. Critically, it also displays a systematic overestimation bias. For predicted scores above 80, a preponderance of residuals are negative, indicating that the model's predictions are consistently higher than the actual scores in this range. In stark contrast, the plot for the G-R model (Figure 9) demonstrates a superior error profile. The residuals are largely homoscedastic, scattered randomly around the zero line with relatively constant variance. Furthermore, there is no evidence of the systematic bias observed in Figure 8. While the overall error pattern is robust, the presence of several large outliers indicates that the model is still capable of making significant individual errors.

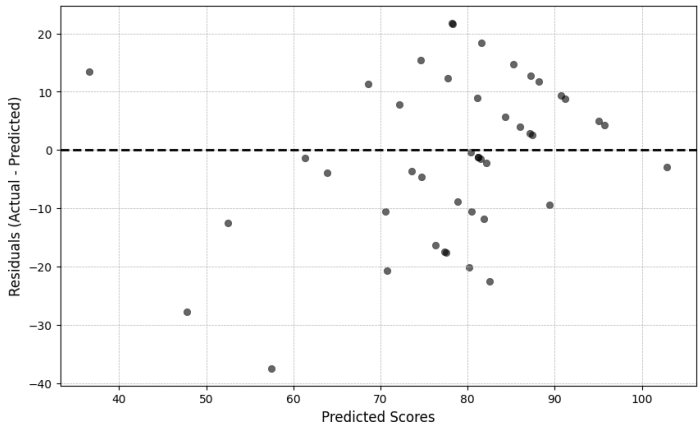

Figure 7. Residual Plot for SVR

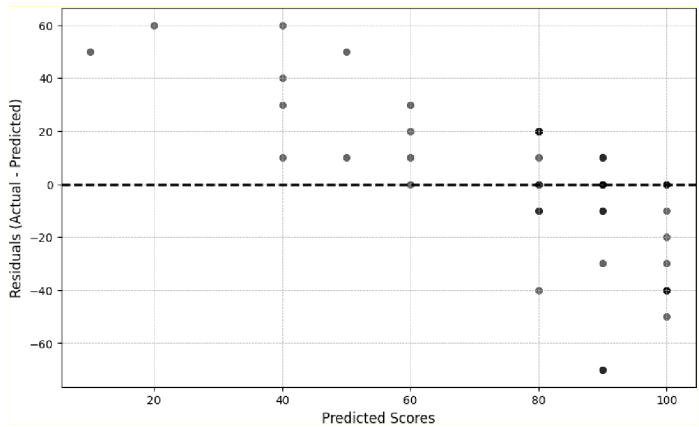

Figure 8. Residual Plot for G

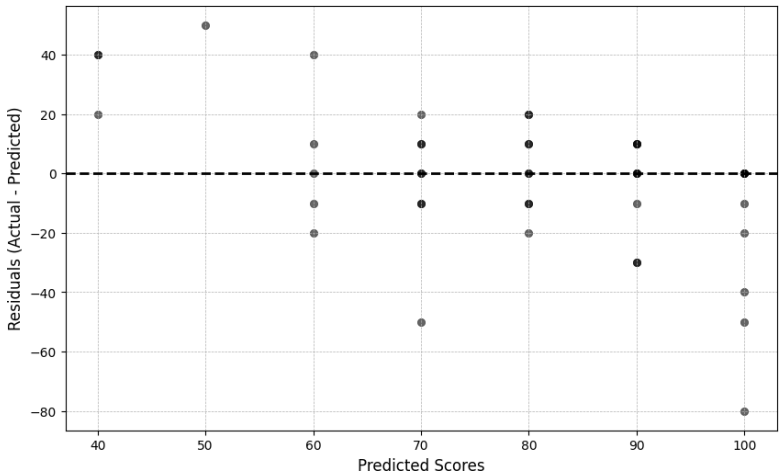

Figure 9. Residual Plot for G-R

The binned confusion matrices, which categorize continuous scores into discrete bins based on the quantiles from Table 1, provide a granular view of these error patterns. The matrix for the SVR model (Figure 10) corroborates its heteroscedastic nature, revealing a "central dumping" phenomenon where the model disproportionately classifies most cases into Bin 1 (70- 80). This strategy results in a near-total failure to correctly identify scores in the higher bins (Bin 2 and Bin 3), confirming that its smaller error variance at high predictions is an artifact of systematic underestimation. The matrix for the G model (Figure 11) provides definitive evidence of the flaws seen in its residual plot. The error pattern is chaotic, with misclassifications occurring between non-adjacent bins. Most importantly, the final column (Predicted Score Bin 3) is composed primarily of instances from lower actual score bins, quantitatively confirming the model's strong systematic overestimation bias. Finally, the matrix for the G-R model (Figure 12) aligns with its well-behaved residual plot. It shows the most balanced performance, with more logical, adjacent-bin errors and a significantly improved ability to correctly classify scores in the higher bins. The remaining cross-category misclassifications correspond to the outliers noted in Figure 9.

The combined visual evidence indicates a clear hierarchy of model quality. The G-R model, which is the proposed model, is demonstrably superior, exhibiting unbiased and largely homoscedastic errors. The SVR model is fundamentally flawed by heteroscedasticity and a simplistic predictive strategy. The G model is the least reliable, suffering from both high error variance and a systematic directional bias.

4.3.2 Analysis of Answer Length Bias

This section investigates the extent to which model predictions are influenced by a common superficial heuristic: the length of the student's answer. The presence of such a bias undermines a model's claim to be assessing content quality. The analysis is based on Pearson and Spearman correlation results presented in Table 6.

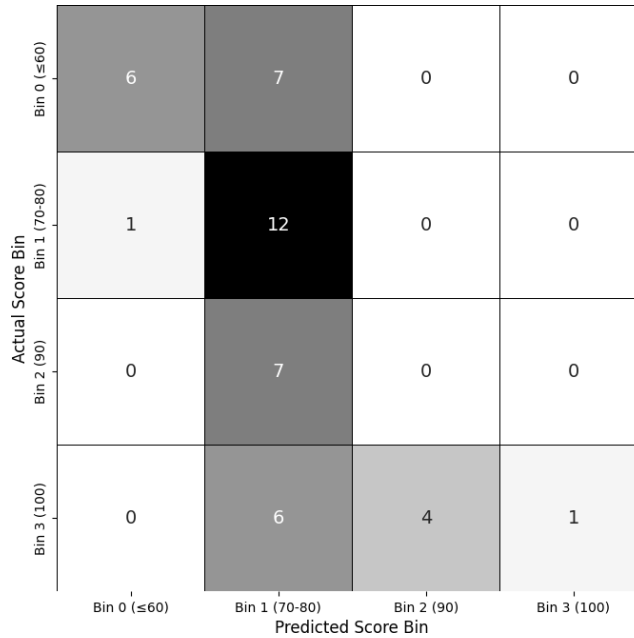

Figure 10. Binned Confusion Matrix for SVR

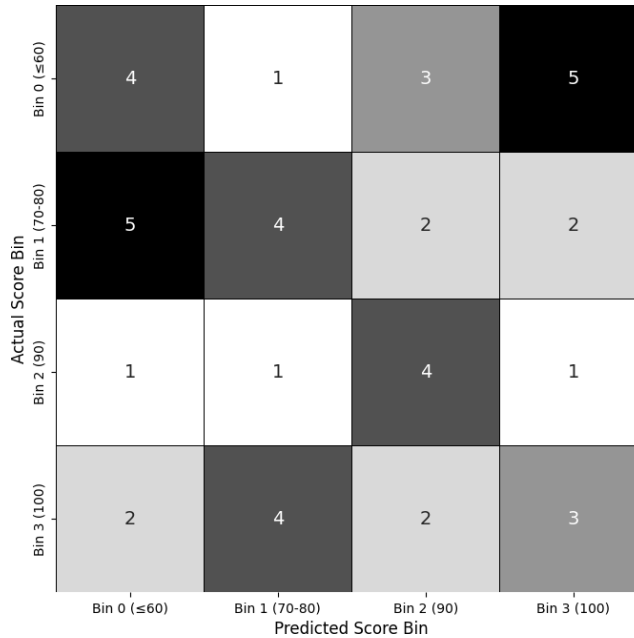

Figure 11. Binned Confusion Matrix for G

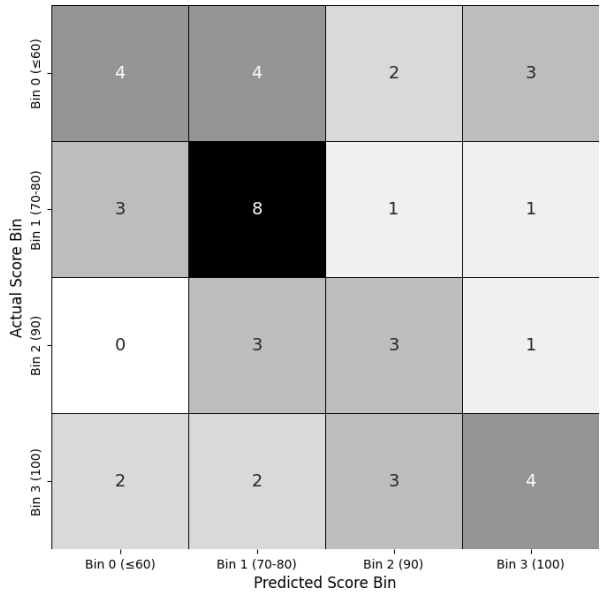

Figure 12. Binned Confusion Matrix for G-R

Table 6. Correlation Analysis of Student's Answer Length vs Predicted Score

Model	Pearson Correlation	Pearson p-value	Spearman Correlation	Spearman p-value
LinearR	0.4008	0.0070	**0.3674**	0.0142
RR	0.6745	**< 0.0001**	0.6755	< 0.0001
LassoR	0.6548	< 0.0001	**0.6664**	< 0.0001
ENR	0.9982	**< 0.0001**	0.9956	< 0.0001
SVR	0.7573	< 0.0001	**0.8289**	< 0.0001
GBR	0.6344	**< 0.0001**	0.6718	< 0.0001
RFR	0.6539	< 0.0001	**0.7086**	< 0.0001
KNR	0.7341	**< 0.0001**	0.7889	< 0.0001
NN	0.9030	< 0.0001	**0.9293**	< 0.0001
G-R	**-0.0024**	**0.9875**	**0.0413**	**0.7901**
G	**-0.2634**	**0.0841**	**-0.3077**	0.0422

The correlation data reveals a critical vulnerability in most of the tested models. A large majority exhibit strong to moderate, statistically significant positive correlations between answer length and predicted score. For the SVR model, the Spearman correlation was exceptionally high ($\rho = 0.8289$, $p < 0.0001$). This indicates these models have learned the fallacious heuristic that longer answers warrant higher scores. The G model displays a statistically significant negative monotonic correlation (Spearman's $\rho = -0.3077$, $p = 0.0422$), indicating it also possesses a length bias, albeit a different one, where it tends to penalize longer answers.

The proposed model (G-R) is the sole exception. It returned correlation coefficients near zero for both Pearson and Spearman metrics, with p-values ($p > 0.79$) indicating a complete lack of statistical significance. This quantitative data provides powerful, independent validation of the findings from the diagnostic plots. The G-R model, identified as the most robust from a qualitative perspective, is also the only model demonstrated to be free of length bias. This analysis demonstrates that even models with acceptable aggregate performance metrics can be fundamentally unreliable if their predictive power stems from superficial heuristics rather than a valid assessment of content. The proposed model is the only one evaluated that meets the criteria for both robust error characteristics and freedom from the tested heuristic bias.

4.4 Qualitative Analysis: Calibration and Superficiality

To gain deeper insights into the behavior of the models beyond aggregate metrics, a qualitative analysis was performed on selected examples from the test set. Table 7 presents two illustrative cases, highlighting instances of calibrated nuance and susceptibility to superficiality.

Table 7. Qualitative Analysis of Model Predictions on Selected Test Samples

Metric / Data	Example 1: Calibrated Nuance (ID 20)	Example 2: Susceptibility to Superficiality (ID 47)
Human Score (Ground Truth)	80.00	20.00
G-R Prediction	80.00	100.00
G Prediction	100.00	90.00
SVR Prediction	68.62	47.76
Professor's Answer	komputer model von neumann model arsitektur dideskripsikan john von neumann tahun 1945 first draft of a report on the edvac model arsitektur dijadikan basis komputer unit pemrosesan berisi alu register prosesor unit kontrol berisi instruction register program counter memori untuk menyimpan data instruksi penyimpanan eksternal mekanisme input output	client server jaringan komputer model komunikasi terjadi dua belah pihak penyedia layanan data disebut server pengguna peminta layanan data disebut client umumnya kedua belah pihak berkomunikasi hardware berbeda terjadi satu mesin sistem
Student's Answer	komputer memiliki processing unit memory input ouput	komputer menggunakan layanan tersedia jaringan

In Example 1, the student's answer ("komputer memiliki processing unit memory input ouput" – a computer has a processing unit, memory, input, output) is short but accurately captures core components of the Von Neumann architecture described in the professor's detailed answer. The human grader assigned a score of 80. The G-R model predicted a score of 80, perfectly matching the human judgment. This suggests that the reasoning-guided model was able to appreciate the conciseness and correctness of the student's response without being penalized for its brevity relative to

the comprehensive professor's answer. In contrast, the G model over-predicted with a score of 100, potentially due to matching key terms without a deeper assessment of completeness. The SVR model under-predicted with a score of 68.62. This case highlights the proposed model's capacity for calibrated nuance, aligning well with human assessment on answers that are correct but not exhaustive.

Example 2 reveals a key vulnerability, particularly in the LLM-based approaches. The student's answer ("komputer menggunakan layanan tersedia jaringan" – computers use available network services) is on-topic in relation to the professor's answer about client-server architecture but is extremely shallow and lacks substantive detail. The human grader assigned a low score of 20, reflecting this lack of depth. However, both LLM variants assigned catastrophically high scores: the G-R model predicted 100, and the G model predicted 90. These inflated scores were likely triggered by the presence of relevant keywords such as "jaringan" (network) and "layanan" (services), without the models adequately discerning the superficial nature of the statement. The SVR model, while still over-predicting at 47.76, was considerably closer to the human score than the LLMs. This example indicates that the LLMs, even in their reasoning-guided fine-tuned state, can struggle to differentiate between deep understanding and superficial keyword matching.

This qualitative analysis underscores that while the reasoning-guided approach demonstrably improves generalization and calibration in many instances (as seen in ID 20 and the overall quantitative metrics), it does not entirely resolve the challenge of assessing the true depth of student understanding. The failure on ID 47 suggests that the model, despite being trained to generate a reason, might still produce reasons based on relatively superficial cues if those cues were correlated with high scores in the training data. The distilled reasoning from the teacher model, or the limited examples in the training set, may not have provided sufficient signal to strongly penalize shallow answers that happen to use correct terminology. This points to a critical area for future improvement: enhancing the model's ability to look beyond surface features to evaluate the substantive content of student responses.

4.5 Implications of Findings

For NLP researchers, the demonstration that "reasoning as regularization" can significantly improve LLM performance and robustness in low-resource domains offers a tangible methodological contribution. The approach of using distilled reasoning from a more capable teacher model as a co-training target for a smaller student model provides a template that could be adapted for other specialized NLP tasks where data scarcity is a bottleneck and nuanced evaluation is required. This study provides empirical evidence for the effectiveness of this technique, particularly in transforming a model prone to memorization into one that exhibits genuine generalization on the target task. The success of using distilled reasoning suggests broader applicability beyond ASAG. This technique could potentially be adapted for other NLP tasks where LLMs suffer from overfitting in low-data scenarios, or where smaller, more efficient models are needed that can still capture some of the nuanced capabilities of larger models. The underlying principle, distilling complex intermediate outputs (like reasoning, planning steps, or structured explanations) from a large teacher model to train

a smaller student model, could serve as a general strategy for creating more robust and efficient specialized LLMs across various domains.

For educators and EdTech developers, this research illuminates a promising, albeit computationally intensive, pathway towards creating more sophisticated, reliable, and potentially feedback-rich automated grading tools, even for under-resourced languages like Indonesian. The high accuracy achieved by the G-R model (QWK 0.7791) suggests its potential for practical application in educational settings, provided the computational costs can be managed. The qualitative analysis, particularly the identification of "susceptibility to superficiality," also offers valuable insights into where human oversight and intervention remain critical, guiding the development of hybrid human-AI grading systems.

4.6 Limitations and Future Work

Despite the promising results, this study has several limitations that warrant discussion and offer directions for future research.

First, the "Susceptibility to Superficiality," as revealed in the qualitative analysis (Example ID 47), is a significant concern. Both LLM variants, including the reasoning-guided model, assigned inappropriately high scores to a shallow answer that contained relevant keywords. This indicates that the models, in their current state, may not adequately distinguish between superficial keyword matching and genuine, deep understanding. Future work should focus on mitigating this vulnerability. One approach could be to enrich the fine-tuning dataset with a more diverse range of answer qualities, specifically including examples of deliberately shallow, cleverly incorrect, or off-topic but keyword-rich answers. Training the model on such adversarial or contrastive examples could teach it a deeper level of discernment.

Second, while this study focused on the instrumental value of reasoning as a powerful regularizer, it did not include a formal quantitative assessment of the intrinsic quality of the rationales generated by the student model during inference. A crucial direction for future research is to conduct such a quantitative human evaluation. This would involve creating a detailed scoring rubric with well-defined criteria, such as Coherence (the logical structure and clarity of the rationale), Correctness (the factual accuracy of its claims), and Sufficiency (whether the provided justification is detailed enough to support the predicted score). To ensure objective and reliable results, this analysis should involve multiple trained raters with domain expertise, and inter-rater reliability scores (e.g., Fleiss' Kappa or Krippendorff's Alpha) should be calculated. This step would provide a direct measure of the model's ability to produce trustworthy explanations and would enable a more granular analysis of its failure modes, which is essential for building confidence in its practical application.

Third, the small dataset size (n=220 total, 176 for training), while intentionally chosen to reflect a low-resource scenario, inherently limits the statistical power and generalizability of the findings. Although the reasoning-guided approach demonstrated clear superiority within this context, further validation on larger and more diverse Indonesian ASAG datasets is a necessary next step. Additionally, exploring the applicability of this methodology to other low-resource languages and different

subject domains would be valuable to confirm the broader relevance of the "reasoning as regularization" principle.

Fourth, the training and test performance curves (Figure 6), even for the reasoning-guided model, exhibited some volatility. While the reasoning component clearly acted as a regularizer, there might be scope for further stabilization and improvement in generalization. Future research could explore the integration of alternative or complementary regularization techniques, such as different optimizers, adjusted dropout rates, or weight decay , in conjunction with the reasoning-guided fine-tuning approach.

Fifth, the effectiveness of the reasoning-guided method is fundamentally bottlenecked by the quality of the distilled reasoning from the teacher model (DeepSeek R1-0528). To mitigate the risk of flawed or biased rationales, every entry in the dataset was manually verified for quality. Despite this curation, the student model's performance ceiling is intrinsically linked to the caliber of these teacher-generated rationales. Any remaining subtle flaws, biases, or superficial cues in the verified reasoning could still be transferred to the student model, limiting the efficacy of the regularization. Future work could investigate the impact of different teacher models or explore methods for further refining the distilled reasoning. More advanced techniques could even empower the student model to critique or improve upon the teacher's rationales. This highlights that the model's ability to differentiate deep understanding from superficial responses is contingent on the signals in its training data, including the distilled reasoning. To achieve finer levels of discernment, future enhancements must therefore focus not only on the reasoning-guided methodology itself but also on elevating the quality and diversity of these core training signals.

Finally, the computational cost of the G-R model, particularly its significantly longer inference time, poses a practical limitation for certain applications. While the improved accuracy may justify the cost in some scenarios, future work could explore methods to optimize the inference efficiency of models that generate extended reasoning chains. This might involve techniques like model distillation (i.e., training an even smaller model to mimic the reasoning-guided model's outputs), quantization beyond 4-bits if feasible, or pruning.

Addressing these limitations will be crucial for advancing the development of robust, reliable, and practical ASAG systems, especially in challenging low-resource environments.

5. Conclusion

The effective application of data-hungry Large Language Models to specialized, low-resource Automated Short Answer Grading (ASAG) tasks, such as for the Indonesian language, presents a formidable challenge. Standard fine-tuning approaches often lead to debilitating overfitting, severely curtailing the practical utility of these powerful models when training data is scarce. This research confronted this challenge directly, investigating a novel fine-tuning strategy to mitigate overfitting and enhance model robustness. We demonstrated that fine-tuning an LLM (Gemma-3-1b-it) with distilled analytical reasoning, sourced from a more capable teacher model, as a co-training target serves as an effective structural regularizer. This approach trans-

formed a model that would otherwise merely memorize the limited training data into one that genuinely generalizes from it. The resulting reasoning-guided model (G-R) achieved state-of-the-art performance on an expanded Indonesian ASAG dataset, significantly outperforming nine traditional machine learning baselines and a standard fine-tuned LLM counterpart.

The primary contributions of this work are threefold. First, we introduce and validate "reasoning as regularization" as a novel, effective method to combat the catastrophic overfitting that typically occurs when fine-tuning LLMs on small, specialized datasets. Second, this approach achieves state-of-the-art performance on an Indonesian ASAG task (QWK 0.7791, Spearman's 0.8276), establishing a new and more robust benchmark for this low-resource domain. Furthermore, robust diagnostic analysis demonstrates the proposed model's superior, homoscedastic error profile and, critically, its freedom from the answer-length bias that systematically compromised every other traditional and LLM-based model tested. Third, we provide a detailed quantitative analysis of the performance-versus-efficiency trade-offs inherent to this approach, revealing the substantial computational costs (over 35× inference time) required to achieve higher accuracy through reasoning generation.

Collectively, these findings provide a methodological template for applying large language models more successfully and responsibly in data-scarce environments. This work paves the way for developing more nuanced, reliable, and fair automated assessment tools in diverse linguistic and educational contexts.

6. Conclusion

The effective application of data-hungry Large Language Models to specialized, low-resource Automated Short Answer Grading (ASAG) tasks, such as for the Indonesian language, presents a formidable challenge. Standard fine-tuning approaches often lead to debilitating overfitting, severely curtailing the practical utility of these powerful models when training data is scarce. This research confronted this challenge directly, investigating a novel fine-tuning strategy to mitigate overfitting and enhance model robustness. We demonstrated that fine-tuning an LLM (Gemma-3-1b-it) with distilled analytical reasoning, sourced from a more capable teacher model, as a co-training target serves as an effective structural regularizer. This approach transformed a model that would otherwise merely memorize the limited training data into one that genuinely generalizes from it. The resulting reasoning-guided model (G-R) achieved state-of-the-art performance on an expanded Indonesian ASAG dataset, significantly outperforming nine traditional machine learning baselines and a standard fine-tuned LLM counterpart.

The primary contributions of this work are threefold. First, we introduce and validate "reasoning as regularization" as a novel, effective method to combat the catastrophic overfitting that typically occurs when fine-tuning LLMs on small, specialized datasets. Second, this approach achieves state-of-the-art performance on an Indonesian ASAG task (QWK 0.7791, Spearman's 0.8276), establishing a new and more robust benchmark for this low-resource domain. Furthermore, robust diagnostic analysis demonstrates the proposed model's superior, homoscedastic error profile and, critically, its freedom from the answer-length bias that systematically compromised

every other traditional and LLM-based model tested. Third, we provide a detailed quantitative analysis of the performance-versus-efficiency trade-offs inherent to this approach, revealing the substantial computational costs (over 35× inference time) required to achieve higher accuracy through reasoning generation.

Collectively, these findings provide a methodological template for applying large language models more successfully and responsibly in data-scarce environments. This work paves the way for developing more nuanced, reliable, and fair automated assessment tools in diverse linguistic and educational contexts.

References

[1] S. Haller et al. "Survey on Automated Short Answer Grading with Deep Learning: from Word Embeddings to Transformers". In: (Mar. 2022). Accessed: Jun. 10, 2025. URL: https://arxiv.org/pdf/2204.03503.

[2] R. Weegar and P. Idestam-Almquist. "Reducing Workload in Short Answer Grading Using Machine Learning". In: *International Journal of Artificial Intelligence in Education* 34.2 (June 2024), pp. 247–273. DOI: 10.1007/S40593-022-00322-1.

[3] L. Anglin et al. "Improving the Efficiency and Effectiveness of Grading Through the Use of Computer-Assisted Grading Rubrics". In: *Decision Sciences Journal of Innovative Education* 6.1 (Jan. 2008), pp. 51–73. DOI: 10.1111/J.1540-4609.2007.00153.X.

[4] F. Chai et al. "Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness". In: *Frontiers in Psychology* 15 (Feb. 2024), p. 1221177. DOI: 10.3389/FPSYG.2024.1221177.

[5] Z. (Helen) Wang, J. Pei, and J. Li. *30 Million Canvas Grading Records Reveal Widespread Sequential Bias and System-Induced Surname Initial Disparity*. Accessed: Jun. 10, 2025. Oct. 2023. URL: https://papers.ssrn.com/abstract=4603146.

[6] J. M. Malouff, A. J. Emmerton, and N. S. Schutte. "The Risk of a Halo Bias as a Reason to Keep Students Anonymous During Grading". In: *Teaching of Psychology* 40.3 (2013), pp. 233–237. DOI: 10.1177/0098628313487425.

[7] J. Klein. "The failure of a decision support system: inconsistency in test grading by teachers". In: *Teaching and Teacher Education* 18.8 (Nov. 2002), pp. 1023–1033. DOI: 10.1016/S0742-051X(02)00057-4.

[8] Ó. Cuéllar, M. Contero, and M. Hincapié. "Personalized and Timely Feedback in Online Education: Enhancing Learning with Deep Learning and Large Language Models". In: *Multimodal Technologies and Interaction* 9.5 (May 2025), p. 45. DOI: 10.3390/MTI9050045.

[9] E. Del Gobbo et al. "GradeAid: a framework for automatic short answers grading in educational contexts-design, implementation and evaluation". In: *Knowledge and Information Systems* 65 (2023), pp. 4295–4334. DOI: 10.1007/s10115-023-01892-9.

[10] C. Zhao, M. Silva, and S. Poulsen. *Language Models are Few-Shot Graders*. Accessed: Jun. 10, 2025. Feb. 2025. URL: https://arxiv.org/pdf/2502.13337.

[11] A. F. Aji et al. "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Vol. 1. Mar. 2022, pp. 7226–7249. DOI: 10.18653/v1/2022.acl-long.500.

[12] S. Cahyawijaya et al. "NusaWrites: Constructing High-Quality Corpora for Underrepresented and Extremely Low-Resource Languages". In: Sept. 2023, pp. 921–945. DOI: 10.18653/v1/2023.ijcnlp-main.60.

[13] L. Susanto et al. "Replicable Benchmarking of Neural Machine Translation (NMT) on Low-Resource Local Languages in Indonesia". In: Nov. 2023, pp. 100–115. DOI: 10.18653/v1/2023.sealp-1.8.

[14] H. Xu et al. *Large Language Models for Education: A Survey*. Accessed: Jun. 10, 2025. May 2024. URL: https://arxiv.org/pdf/2405.13001.

[15] K. Tirumala et al. *Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models*. Accessed: Jun. 10, 2025. May 2022. URL: https://arxiv.org/pdf/2205.10770.

[16] D. Hernandez et al. *Scaling Laws and Interpretability of Learning from Repeated Data*. Accessed: Jun. 10, 2025. May 2022. URL: https://arxiv.org/pdf/2205.10487.

[17] S. A. Mahmood and M. A. Abdulsamad. "Automatic assessment of short answer questions: Review". In: *Edelweiss Applied Science and Technology* 8.6 (2024). Accessed: Jun. 10, 2025, pp. 9158–9176. URL: https://ideas.repec.org/a/ajp/edwast/v8y2024i6p9158-9176id3956.html.

[18] N. LaVoie et al. "Using Latent Semantic Analysis to Score Short Answer Constructed Responses: Automated Scoring of the Consequences Test". In: *Educational and Psychological Measurement* 80.2 (Apr. 2019), p. 399. DOI: 10.1177/0013164419860575.

[19] N. Reimers and I. Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *EMNLP-IJCNLP 2019 - Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*. Aug. 2019, pp. 3982–3992. DOI: 10.18653/v1/d19-1410.

[20] *SentenceTransformers Documentation — Sentence Transformers documentation*. Accessed: Jun. 10, 2025. URL: https://sbert.net/.

[21] J. Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL HLT 2019 - Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1. Accessed: Jun. 10, 2025. Oct. 2018, pp. 4171–4186. URL: https://arxiv.org/pdf/1810.04805.

[22] OpenAI et al. *GPT-4 Technical Report*. Accessed: Jun. 10, 2025. Mar. 2023. URL: https://arxiv.org/pdf/2303.08774.

[23] G. Team et al. *Gemma 3 Technical Report*. Accessed: Jun. 10, 2025. Mar. 2025. URL: https://arxiv.org/pdf/2503.19786.

[24] Z. Han et al. *Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey*. Accessed: Jun. 10, 2025. Mar. 2024. URL: https://arxiv.org/pdf/2403.14608.

[25] E. Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models". In: *ICLR 2022 - 10th International Conference on Learning Representations*. Accessed: Jun. 10, 2025. June 2021. URL: https://arxiv.org/pdf/2106.09685.

[26] J. Wei et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems*. Vol. 35. Accessed: Jun. 10, 2025. Jan. 2022. URL: https://arxiv.org/pdf/2201.11903.

[27] *deepseek-ai/DeepSeek-R1-0528 · Hugging Face*. Accessed: Jun. 10, 2025. URL: https://huggingface.co/deepseek-ai/DeepSeek-R1-0528.

[28] F. Koto et al. *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP*. Nov. 2020.

[29] W. Wongso et al. *NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural*. Accessed: Jun. 10, 2025. Mar. 2024. URL: https://arxiv.org/pdf/2403.01817.

[30] W. Puspitasari, D. Ramdani, and A. M. Maulana. "IndoT5 (Text-to-Text Transfer Transformer) Algorithm for Paraphrasing Indonesian Language Islamic Sermon Manuscripts". In: *Khazanah Journal of Religion and Technology* 2.2 (Jan. 2024), pp. 63–73. DOI: 10.15575/KJRT.V2I2.1093.

[31] A. A. Putri Ratna et al. "Automatic Essay Grading for Bahasa Indonesia with Support Vector Machine and Latent Semantic Analysis". In: *ICECOS 2019 - 3rd International Conference on Electrical Engineering and Computer Science*. Oct. 2019, pp. 363–367. DOI: 10.1109/ICECOS47637.2019.8984528.

[32] M. C. Wijaya. "Automatic Short Answer Grading System in Indonesian Language Using BERT Machine Learning". In: *Revue d'Intelligence Artificielle* 35.6 (Dec. 2021), pp. 503–509. DOI: 10.18280/RIA.350609.

[33] M. Kharis, K. Laksono, and Suhartono. "Utilization of NLP-Technology in Current Applications for Education and Research by Indonesian Student, Teacher, and Lecturer". In: *Journal of Higher Education Theory and Practice* 22.14 (Nov. 2022), pp. 170–178. DOI: 10.33423/JHETP.V22I14.5544.