FeedEval: Pedagogically Aligned Evaluation of LLM-Generated Essay Feedback

Seongyeub Chu[1], Jongwoo Kim[2], Mun Yong Yi[1]*
[1]Graduate School of Data Science, KAIST
[2]Department of Industrial & Systems Engineering, KAIST
{chseye7, gsds4885, munyi}@kaist.ac.kr

Abstract

Going beyond the prediction of numerical scores, recent research in automated essay scoring has increasingly emphasized the generation of high-quality feedback that provides justification and actionable guidance. To mitigate the high cost of expert annotation, prior work has commonly relied on LLM-generated feedback to train essay assessment models. However, such feedback is often incorporated without explicit quality validation, resulting in the propagation of noise in downstream applications. To address this limitation, we propose FeedEval, an LLM-based framework for evaluating LLM-generated essay feedback along three pedagogically grounded dimensions: specificity, helpfulness, and validity. FeedEval employs dimension-specialized LLM evaluators trained on datasets curated in this study to assess multiple feedback candidates and select high-quality feedback for downstream use. Experiments on the ASAP++ benchmark show that FeedEval closely aligns with human expert judgments and that essay scoring models trained with FeedEval-filtered high-quality feedback achieve superior scoring performance. Furthermore, revision experiments using small LLMs show that the high-quality feedback identified by FeedEval leads to more effective essay revisions. We release our code and curated datasets at: **https://github.com/BBeeChu/FeedEval.git**.

1 Introduction

Automated essay assessment has evolved from feature-engineered approaches to pre-trained and large language models (LLMs) (Ramesh and Sanampudi, 2022; Li and Ng, 2024; Misgna et al., 2024). While early work focused primarily on essay scoring, recent studies have explored joint modeling of scoring and feedback generation to provide pedagogically meaningful guidance beyond numerical scores. Developing essay feedback generation models typically requires pedagogically grounded feedback annotated by domain experts, which is costly and impractical in real-world settings (Macina et al., 2025; Li and Ng, 2024). To address this limitation, recent approaches increasingly rely on LLM-driven synthetic data generation for essay feedback (Li et al., 2023; Do et al., 2025).

In essay scoring with feedback generation, the quality of feedback labels is crucial for training models to predict scores accurately and generate pedagogically useful feedback. However, prior approaches use LLM-generated feedback as labels without explicit quality validation. As shown in Figure 1 (Feedback 2), using LLM-generated feedback without quality evaluation can introduce low-quality outputs that do not refer to the content of the essay, deviate from the scoring rubric, or provide limited actionable guidance, ultimately degrading downstream applications such as model training. To address this issue, we propose **FeedEval** (**Feed**back **Eval**uation), an LLM-based framework for evaluating the pedagogical quality of multiple LLM-generated essay feedback candidates and identifying high-quality feedback.

Through FeedEval, we fine-tune LLMs to assess feedback quality along three **pedagogically grounded dimensions** — specificity (Hattie and Timperley, 2007; Shute, 2008), helpfulness (Steiss et al., 2024), and validity (Black and Wiliam, 2009). This design yields evaluations that closely align with human teachers' judgments of feedback quality, which we refer to as **pedagogically aligned evaluation** and confirm through multiple experiments involving experts in the educational domain.

As shown in Figure 1 (Feedback 1), FeedEval filters high-quality feedback closely tailored to both the essay content and the scoring rubric, which can further serve as reliable supervision for downstream tasks, including essay evaluation model training. Specifically, we construct or adapt

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 12648–12674

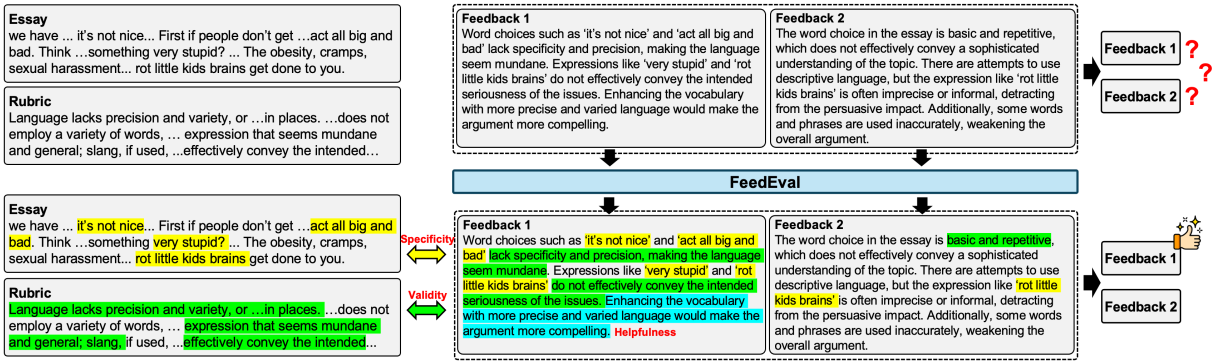

Figure 1: FeedEval evaluates the quality of multiple feedback candidates for the same essay by assessing how well they reference the essay, align with the rubric, and provide actionable revision suggestions.

dimension-relevant datasets to train the LLM evaluators for each dimension. Then, we prompt an LLM to generate multiple feedback candidates for each essay via temperature sampling and evaluate them using the dimension-specific LLM evaluators. Based on the resulting scores, FeedEval selects high-quality feedback from multiple candidates.

We conduct our experiments on the ASAP++ dataset of student essays with human-annotated multi-trait scores. We first validate the alignment of FeedEval with educational experts in judging the quality of LLM-generated essay feedback. We then evaluate the effectiveness of FeedEval by comparing the essay scoring accuracy of LLMs trained on high- and low-quality feedback filtered by FeedEval. In addition, we examine the pedagogical usefulness of the high-quality feedback through essay revision experiments and human evaluations. Our extensive experiments show that FeedEval achieves close alignment with human expert judgments, that training with FeedEval-filtered high-quality feedback leads to more accurate essay scoring, and that the resulting feedback is more pedagogically helpful than its low-quality counterpart.

To summarize, our main contributions are as follows: (1) we propose **FeedEval**, an LLM-based framework that evaluates essay feedback along three pedagogical dimensions—specificity, helpfulness, and validity—and release **SpecEval**, the first dataset for training essay specificity evaluation models, (2) we demonstrate that FeedEval exhibits close agreement with expert judgments in essay feedback evaluation via human expert evaluations, and (3) through extensive experiments, we show that FeedEval-filtered high-quality feedback leads to more accurate essay scoring and is pedagogically more meaningful than low-quality feedback.

2 Related Work

2.1 LLM-based Essay Assessment

Recent research on automated essay assessment has increasingly focused on leveraging LLMs, including conventional language models (e.g., BERT), for essay scoring and feedback generation (Yang et al., 2020; Wang et al., 2022; Do et al., 2023, 2024; Lee et al., 2024; Li and Ng, 2025). However, jointly generating essay scores and feedback typically requires expert-written, pedagogically grounded feedback, which is costly and impractical in real-world settings (Li and Ng, 2024). To address this limitation, recent approaches use LLMs to generate essay feedback or rationales (Chu et al., 2025b; Do et al., 2025; Li et al., 2023), and leverage them together with human-annotated scores as supervision for training essay assessment models.

Despite these advances, most existing methods implicitly assume that LLM-generated feedback is pedagogically reliable and do not explicitly assess its quality. As a result, low-quality feedback (e.g., overly generic or misleading feedback) may be included in the training data, potentially degrading model performance. In contrast, our work systematically evaluates the pedagogical quality of LLM-generated feedback and filters high-quality feedback, which can be used to develop more reliable LLM-based essay assessment models.

2.2 Evaluation of LLM-generated Feedback

Evaluating the quality of educational feedback is a longstanding challenge in the learning sciences area (Stahl et al., 2024; Meyer et al., 2024; Behzad et al., 2024a). Prior work emphasizes that effective feedback should be grounded in students' work, rubric-aligned, and provide actionable guidance (Hattie and Timperley, 2007; Shute, 2008;

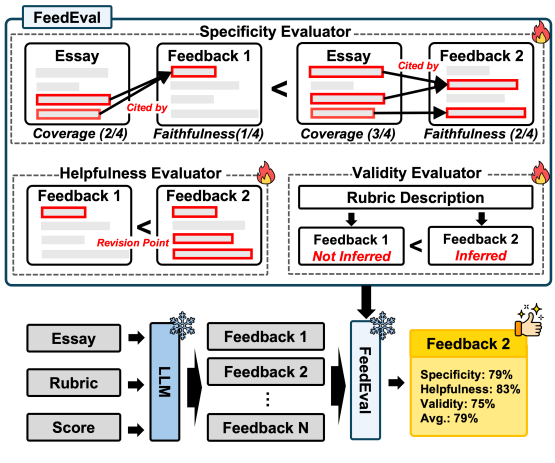

Figure 2: Overview of the proposed FeedEval framework. FeedEval consists of three evaluators, each trained on a dataset corresponding to a specific evaluation dimension. Given multiple feedback candidates generated by an LLM, FeedEval evaluates their quality along three dimensions and selects the highest-quality feedback.

Black and Wiliam, 2009; Steiss et al., 2024). However, these criteria are inherently open-ended and context-dependent, requiring expert pedagogical judgment, which makes high-quality feedback annotation costly, labor-intensive, and hard to scale (Macina et al., 2025; Li and Ng, 2024). As a result, reference-based metrics such as BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2019) are ill-suited for evaluating the quality of LLM-generated feedback, while LLM-as-a-judge approaches show limited agreement with human experts and raise pedagogical validity concerns (Li et al., 2025; Liang et al., 2024; Maurya et al., 2025).

To address this gap, we propose FeedEval, a feedback evaluation framework tailored to essay assessment. FeedEval provides a scalable, pedagogy-aware proxy for evaluating and filtering LLM-generated feedback, and is validated through direct comparisons with expert teachers' judgments.

3 FeedEval Framework

We propose a new framework, FeedEval, for evaluating LLM-generated essay feedback across multiple traits and filtering high-quality feedback from multiple candidates. Figure 2 illustrates FeedEval in detail. Its components are explained below.

3.1 Evaluation Dimensions

Extant literature on educational feedback underscores several multifaceted components essential for instructional efficacy. Specifically, effective feedback must facilitate feed-up by clarifying learning objectives, feed-back through precise diagnostic assessments of current performance, and feed-forward by offering strategic pathways for improvement (Hattie and Timperley, 2007). Furthermore, integrating the five quality dimensions established in prior research (Scarlatos et al., 2024; Steiss et al., 2024; Black and Wiliam, 2009), feedback should (1) maintain non-revelatory guidance to preserve cognitive challenge (revealing), (2) ensure factual accuracy for reliable support (correctness), (3) provide diagnostic precision in evaluating learner levels (diagnosticity), (4) offer actionable scaffolding for revision (guidance), and (5) employ an encouraging tone to foster motivation (encouragement).

Synthesizing these theoretical foundations, this study evaluates LLM-generated essay feedback across three pedagogically grounded dimensions:

- **Specificity**: Feedback includes explicit references to relevant parts of the student's essay.
- **Helpfulness**: Feedback provides actionable guidance that supports the student's improvement.
- **Validity**: Feedback accurately diagnoses the quality of the student's essay based on the rubric score descriptions.

Regarding the feedback dimensions proposed in Hattie and Timperley 2007, we explicitly omit the feed-up component. In essay scoring, learning targets are inherently defined by the scoring rubrics and remain invariant across samples; thus, reiterating these objectives offers marginal instructional value while potentially inflating feedback length and compromising the efficiency of LLM generation. Instead, our framework concentrates on feedback—operationalized through specificity and validity—and feed-forward, captured by helpfulness, to provide targeted diagnostic and improvement-oriented signals. Furthermore, we refine the five feedback quality dimensions established in prior work (Scarlatos et al., 2024; Steiss et al., 2024; Black and Wiliam, 2009) by excluding the revealing aspect, as essay tasks lack a singular, predefined correct answer. The remaining four attributes are systematically mapped onto our evaluation criteria: correctness and diagnosticity correspond to specificity and validity, respectively, while guidance and encouragement are integrated into helpfulness, as both are essential for scaffolding the learner's iterative refinement process.

Algorithm 1 FeedEval-based Feedback Selection

Input: Essay e; Traits $\mathcal{T}$; Feedback candidates $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ with $\mathcal{F}_t = \{f_t^{(1)}, \ldots, f_t^{(N)}\}$; Rubric score descriptions $\{r_t\}_{t\in\mathcal{T}}$
Output: Selected feedback $\mathcal{F}^\star$
$\mathcal{F}^\star \leftarrow \emptyset$
for $t \in \mathcal{T}$ **do**
  for $i = 1$ to N **do**
    $s_{\text{spec}}^{(i)} \leftarrow \text{Spec}(e, f_t^{(i)})$
    $s_{\text{help}}^{(i)} \leftarrow \text{Help}(e, f_t^{(i)})$
    $s_{\text{valid}}^{(i)} \leftarrow \text{Valid}(r_t, f_t^{(i)})$
  end for
  $\tilde{\mathbf{s}}_{\text{spec}} \leftarrow \text{Softmax}\left(\{s_{\text{spec}}^{(i)}\}_{i=1}^N\right)$
  $\tilde{\mathbf{s}}_{\text{help}} \leftarrow \text{Softmax}\left(\{s_{\text{help}}^{(i)}\}_{i=1}^N\right)$
  $\tilde{\mathbf{s}}_{\text{valid}} \leftarrow \text{Softmax}\left(\{s_{\text{valid}}^{(i)}\}_{i=1}^N\right)$
  $i^\star \leftarrow \arg\max_{i\in\{1,\ldots,N\}} \frac{1}{3}\left(\tilde{\mathbf{s}}_{\text{spec}}^{(i)} + \tilde{\mathbf{s}}_{\text{help}}^{(i)} + \tilde{\mathbf{s}}_{\text{valid}}^{(i)}\right)$
  $\mathcal{F}^\star \leftarrow \mathcal{F}^\star \cup \{f_t^{(i^\star)}\}$
end for
return $\mathcal{F}^\star$

3.2 Evaluation Pipeline

We employ dimension-specific LLM evaluators trained to assess each feedback dimension. As shown in Algorithm 1, FeedEval evaluates multiple feedback candidates across traits and selects the highest-quality one per trait. For each dimension, scores are normalized across feedback candidates using a softmax function. The following sections describe how each evaluator is constructed.

3.2.1 Specificity Evaluator

The specificity evaluator measures how faithfully and widely feedback references a student's essay. Given that no existing dataset explicitly targets the specificity of essay feedback, we construct a new 41K dataset (**SpecEval**[1]) using GPT-4o, leveraging its strong capabilities in document comparison and fine-grained textual alignment (Chu et al., 2025a). For each essay, we generate three feedback variants using different prompt designs (see Appendix A.1), resulting in three essay–feedback pairs. Given each pair, the LLM extracts essay segments directly quoted in the feedback. We then compute two metrics: (1) the proportion of feedback sentences that reference the essay (faithfulness) and (2) the proportion of essay sentences referenced by the feedback (coverage). Specificity is defined as the F1 score of these two metrics. Using these scores, we construct a chosen–rejected pairwise dataset by ranking feedback candidates for the same essay.

To build the specificity evaluator for selecting the superior feedback from a pair, we train an LLM-based reward model on the SpecEval dataset using a binary ranking loss following Ouyang et al. (2022):

$$\mathcal{L}_{\text{rank}} = -\frac{1}{N}\sum_{i=1}^{N} \log \sigma\big(r_\theta(e_i, f_i^+) - r_\theta(e_i, f_i^-) - m\big) \tag{1}$$

where $r_\theta(e, f)$ is the scalar score for an essay e and feedback f, f^+ and f^- denote the chosen and rejected feedback, respectively, m enforces a margin between them (Macina et al., 2025), and σ denotes the sigmoid function.

3.3 Helpfulness Evaluator

The helpfulness evaluator assesses how well an LLM-generated feedback provides actionable revision points—that is, concrete suggestions that a student can directly apply to improve an essay.

To capture this notion, we construct a 14K pairwise dataset by adapting prior feedback datasets (Han et al., 2024; Seo et al., 2025) and human-written feedback samples from the ASAP++ dataset to our task setting. Specifically, we reformat these datasets into chosen–rejected pairs based on whether the feedback offers clear, actionable guidance for revision, and use them to train the helpfulness evaluator (see Appendix B.1 for details of the datasets and reformatting strategy). Following the same training strategy as the specificity evaluator, we train an LLM-based reward model using a binary ranking loss (Ouyang et al., 2022).

3.4 Validity Evaluator

The validity evaluator assesses whether feedback accurately diagnoses a student's essay with respect to the rubric score descriptions. We formulate validity evaluation as a natural language inference (NLI) task, assuming that high-validity feedback should be readily inferable from the rubric score descriptions corresponding to an essay's score. More specifically, we treat the rubric score descriptions as the premise and the feedback as the hypothesis. To train the validity evaluator, we use the Prometheus dataset[2], designed to train LLMs that align closely with human judgments for evaluating open-ended responses based on rubric guidelines. The dataset is a synthetic dataset generated by GPT-4 including open-ended responses, scoring rubrics,

[1] We release the dataset for future research.

[2] https://huggingface.co/datasets/prometheus-eval/Feedback-Collection

scores, and corresponding feedback. We reformulate this dataset into a validity-focused NLI task by pairing rubric score descriptions with feedback: for each response, the rubric score descriptions corresponding to the evaluated score is treated as the premise and the feedback as the hypothesis, labeled as *entailment*, while pairing the same feedback with rubric descriptions from a randomly selected different score level is labeled as *contradiction*. This process yields a 99K paired NLI-style dataset, enabling the model to learn whether feedback accurately reflects the rubric-defined level of an essay.

We train an LLM using the loss:

$$\mathcal{L}_{\text{NLI}} = -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta\left(y_i \mid rubric_i^{pre}, feedback_i^{hyp}\right) \tag{2}$$

where $p_\theta(y \mid rubric^{pre}, feedback^{hyp})$ denotes the probability of generating the NLI label $y \in \{entailment, contradiction\}$, conditioned on the $rubric^{pre}$ and the $feedback^{hyp}$. The probability of generating *entailment* is used as the initial score prior to normalization.

4 Experimental Setup

We conduct extensive experiments to examine (1) the alignment between FeedEval instantiated with different LLMs and human experts in evaluating the quality of essay feedback, (2) the impact of FeedEval-assessed feedback quality on training LLMs for essay scoring, and (3) the effectiveness of FeedEval-assessed feedback quality in guiding essay revision by small LLMs. Our study is guided by the following three research questions:

- **RQ1.** How well does FeedEval, instantiated with different LLMs, align with human expert judgments when evaluating essay feedback across specificity, helpfulness, and validity?
- **RQ2.** How does feedback quality assessed by FeedEval influence the training of LLMs for essay scoring?
- **RQ3.** How does feedback quality assessed by FeedEval affect essay revision by small LLMs?

4.1 Datasets

To evaluate FeedEval's ability to assess essay feedback quality, we use the ASAP++ dataset (Mathias and Bhattacharyya, 2018), an enhanced version of the ASAP dataset (Ben et al., 2012), that provides human-annotated multi-trait scores for English essays across six prompts. As shown in Table 1, different prompts are evaluated using different writing traits. We exclude the original ASAP dataset because it reports only aggregated scores from two annotators, making it difficult to align scores with explicit rubric descriptions.

Dataset	Prompt	# Essays	Traits
ASAP++	1	1783	Over, Cont, WC, Org, SF, Conv
	2	1800	Over, Cont, WC, Org, SF, Conv
	3	1726	Over, Cont, PA, Nar, Lang
	4	1772	Over, Cont, PA, Nar, Lang
	5	1805	Over, Cont, PA, Nar, Lang
	6	1800	Over, Cont, PA, Nar, Lang

Table 1: Prompt-trait composition of the ASAP++ dataset. Traits include Overall (Over), Content (Cont), Word Choice (WC), Organization (Org), Sentence Fluency (SF), Conventions (Conv), Prompt Adherence (PA), Narrativity (Nar), and Language (Lang).

4.2 Human Expert Alignment of FeedEval

We evaluate the alignment between FeedEval and educational experts in judging feedback quality across three dimensions. FeedEval is instantiated with 3B-scale LLM backbones (approximately 3–4B parameters) and fine-tuned on dimension-specific datasets to analyze expert alignment. For comparison, we also include GPT-5.1 and Gemini-2.5-Pro using LLM-as-a-judge prompting, representing widely adopted LLM-based evaluation approaches (Li et al., 2025; Zheng et al., 2023). We employ four prompting strategies: zero-shot, zero-shot with Chain-of-Thought (CoT), few-shot, and few-shot with CoT. Detailed prompt templates are provided in Appendix A.2.

Alignment is assessed via pairwise comparisons constructed from GPT-5.1-generated feedback[3]. Following prior work (Li et al., 2023; Do et al., 2025; Chu et al., 2025b), we design three prompting settings that vary the inclusion of essays, prompts, excerpts, human-annotated scores, and rubric descriptions (see Appendix A.1). Three educational experts selected the better feedback per dimension, and alignment is measured by whether the feedback with the higher FeedEval score matches expert preferences[4]. We report accuracy and F1 scores for the resulting pairwise rankings.

[3] A different model version is used for evaluation to avoid overlap with GPT-4o, which was used to construct the specificity dataset (SpecEval).

[4] Human evaluation details are provided in Appendix C.1.

4.3 Feedback Quality Evaluation

To examine the feedback quality evaluated by FeedEval, we prompt GPT-5.1 to generate eight feedback candidates per essay and trait using temperature sampling (temperature = 0.7). We then select feedback with the highest or lowest average FeedEval scores across specificity, helpfulness, and validity, referred to as high- and low-quality feedback, respectively. The Overall trait is excluded due to the absence of rubric descriptions. To further assess FeedEval's filtering effectiveness, we compare these results with feedback filtered by GPT-5.1, which selects feedback of the highest- or lowest-quality from the candidate set without relying on FeedEval scores. We adopt few-shot prompting with CoT as it demonstrates the highest alignment with human experts among the evaluated proprietary LLM configurations (see Table 2).

4.4 Impact of Feedback Quality on Essay Scoring

Since high-quality LLM-generated rationales effectively supervise smaller LLMs' reasoning (Kang et al., 2023), we train 8B-scale LLMs to jointly generate multi-trait scores and feedback[5]. Outputs follow a structured JSON format, where each trait contains a score and feedback (e.g., `{content: {score:3.0, feedback:...}, word choice: {score:2.0, feedback:...},...}`), while feedback for the Overall trait is set to "NAN". This design supports generating long outputs and enables score prediction followed by feedback generation. Scoring performance is evaluated using quadratic weighted kappa (QWK) (Cohen, 1968) with 5-fold cross-validation, comparing models trained with high- and low-quality feedback labels.

4.5 Impact of Feedback Quality on Essay Revision

We investigate the impact of feedback quality on essay revision through two approaches. First, following prior work (Nair et al., 2024; Dinucu-Jianu et al., 2025), we use small-sized LLMs as student simulators to revise human-written essays guided by feedback of varying quality, and measure revision gains using a fine-tuned essay scoring model. This design is motivated by evidence that small LLMs struggle to produce high-quality text compared to larger models (Song et al., 2025; Eldan and Li, 2023), making them a reasonable proxy for learners who benefit from feedback. Although validation with real students is ultimately necessary, this controlled setting enables scalable and reproducible evaluation of feedback effectiveness without the ethical and privacy constraints associated with studies involving human learners (Macina et al., 2025). Second, we conduct human evaluations in which domain experts compare feedback identified as high- or low-quality by FeedEval to assess its pedagogical usefulness.

[5]Implementation details are provided in Appendix D.

5 Experimental Results

5.1 Human Expert Alignment of FeedEval (RQ1)

5.1.1 Alignment Across Different LLMs

Model	Specificity		Helpfulness		Validity	
	Acc.	F1	Acc.	F1	Acc.	F1
GPT (zero-shot)	0.729	0.833	0.584	0.697	0.640	0.445
GPT (zero-shot+CoT)	0.735	0.831	0.590	0.699	0.656	0.457
GPT (few-shot)	0.742	0.847	0.599	0.706	0.661	0.462
GPT (few-shot+CoT)	0.751	0.859	0.615	0.720	0.682	0.479
Gemini (zero-shot)	0.757	0.845	0.556	0.622	0.613	0.417
Gemini (zero-shot+CoT)	0.761	0.859	0.567	0.639	0.624	0.428
Gemini (few-shot)	0.769	0.856	0.571	0.643	0.638	0.439
Gemini (few-shot+CoT)	0.783	0.869	0.597	0.659	0.653	0.458
Llama3-3B-Inst. (3B-scale)	0.820	0.880	0.864	0.912	**0.835**	**0.709**
Qwen2-3B-Inst. (3B-scale)	0.807	0.870	0.755	0.824	0.833	0.703
Phi-3-Mini (3B-scale)	0.811	0.860	**0.871**	**0.920**	0.820	0.700
Gemma3-Inst. (3B-scale)	**0.832**	**0.893**	0.853	0.895	0.822	0.702

Table 2: Human alignment of FeedEval across three feedback dimensions (pairwise Acc./F1). All 3B-scale models are fine-tuned. GPT and Gemini refer to GPT-5.1 and Gemini-2.5-Pro respectively. The best performances are shown in **bold**, and the second-best are underlined.

Table 2 reports the alignment between FeedEval and human experts across the three feedback quality dimensions for different LLM backbones. **3B-scale models trained on dimension-specific datasets achieve the strong alignment, consistently attaining accuracies above 75% and F1 scores above 70% across the three dimensions**. These fine-tuned models outperform larger frozen models, including GPT-5.1 and Gemini-2.5-Pro, highlighting the effectiveness of dimension-specific fine-tuning for feedback quality evaluation. Among the frozen proprietary models, GPT-5.1—when configured with few-shot prompting and CoT—demonstrates the highest overall alignment with human expert annotations.

Based on its consistent strong performance across all of the dimensions, we adopt Llama3-3B-Instruct as the final FeedEval backbone. Cohen's Kappa of the backbone shows substantial agreement for helpfulness (0.62) and moderate agree-

ment for specificity[6] (0.52) and validity (0.59).

To investigate potential model-dependency and biases stemming from the exclusive use of the GPT family (GPT-4o for SpecEval construction and GPT-5.1 for essay feedback generation), we additionally construct an alternative SpecEval dataset using Llama3-70B. We subsequently train evaluators to compare its alignment with expert annotations as well as with the evaluators trained on GPT-4o-based SpecEval. As a result, we observe that the evaluator trained on the Llama3-70B-synthesized data demonstrate comparable human–LLM agreement in terms of accuracy and F1 score to its GPT-4o-based counterpart. Furthermore, the evaluators trained on the GPT-4o-generated and Llama3-70B-generated SpecEval datasets achieve high agreement scores (exceeding 0.8 in both metrics). These results suggest that both models capture consistent pedagogical characteristics of essay feedback, implying that the choice of generator LLM does not significantly bias the resulting evaluator's performance. Detailed results are shown in Appendix E.

5.1.2 Impact of Feedback-related Knowledge

Knowledge Configuration		Accuracy/F1-score		
Task-related	Feedback-related	Specificity	Helpfulness	Validity
✗	✗	0.404 / 0.489	0.262 / 0.000	0.513 / 0.282
✓	✗	0.644 / 0.748	0.656 / 0.768	0.671 / 0.417
✓	✓	0.820 / 0.880	0.864 / 0.912	0.835 / 0.709

Table 3: Alignment of Llama-based fine-tuned models with expert judgments across knowledge configuration.

Given our goal of enabling LLMs to evaluate feedback quality, we examine the role of feedback-specific knowledge in aligning LLM judgments with those of human experts. To this end, we compare three settings: no fine-tuning, fine-tuning on generic task-oriented data, and fine-tuning on feedback-specific data constructed in this study. For the generic task-oriented fine-tuning, we train specificity and helpfulness evaluators using a human-preference reward dataset[7], and train the validity evaluator on the MNLI dataset[8], formulated as an NLI task. As shown in Table 3, **while the generic task-oriented fine-tuning improves alignment, fine-tuning with our curated datasets yields further gains, demonstrating the importance of feedback-specific knowledge for reliable LLM-based feedback assessment**.

5.1.3 Analysis of LLM-generated Feedback

Score	Rubric	Specificity	Helpfulness	Validity
✓	✓	**56.68**	<u>40.71</u>	**63.72**
✓	✗	2.22	9.48	<u>27.31</u>
✗	✓	<u>41.10</u>	**49.81**	8.97

Table 4: Proportion (%) of feedback-generation methods that achieve the highest score per dimension for each essay. The best results are shown in **bold**, and the second-best are <u>underlined</u>.

We analyze feedback quality across three prompting strategies specified in Section 4.2 by measuring how often each method achieves the highest FeedEval score per essay across specificity, helpfulness, and validity. As shown in Table 4, prompting with both human-annotated scores and rubric score descriptions most consistently yields high-quality feedback, particularly for specificity and validity, underscoring the importance of score–rubric integration. Rubric-only prompting achieves the best performance for helpfulness and the second-best for specificity, highlighting the crucial role of rubric information. Accordingly, we adopt the score–rubric prompting strategy to generate feedback candidates in subsequent experiments.

5.2 Impact of Feedback Quality on Essay Scoring (RQ2)

5.2.1 Essay Scoring Performance

Table 5 reports the essay scoring performance of Llama3-8B-Instruct and Qwen3-8B under different label configurations[9]. The results show that models trained on high-quality feedback filtered by FeedEval consistently outperform those trained on low-quality feedback across all traits. In contrast, models trained on high-quality feedback filtered by GPT-5.1 do not yield consistent performance gains over their low-quality counterparts, **highlighting FeedEval's superior ability to assess and filter pedagogically helpful feedback and the effectiveness of high-quality feedback filtered by FeedEval as supervision for essay scoring**. Moreover, models trained on FeedEval-selected high-quality feedback consistently outperform those trained on

[6]The GPT-4o-based specificity computation used for SpecEval dataset construction shows high human alignment (accuracy: 89.2%, F1: 92.0%, Cohen's Kappa: 0.72); however, we use the fine-tuned 3B model for efficiency.

[7]Anthropic RLHF (Bai et al., 2022), https://huggingface.co/datasets/Anthropic/hh-rlhf.

[8]MNLI (Williams et al., 2018), https://huggingface.co/datasets/nyu-mll/multi_nli.

[9]Additional experiments on another dataset to examine the generalizability of FeedEval's impact on essay feedback quality evaluation are provided in the Appendix G.

LLM	Assessment	Feedback Quality	Traits (Prediction Order: ←) Over	Cont	PA	Lang	Nar	Org	Conv	WC	SF	Avg↑ (SD↓)
Llama3-8B-Inst.	Score Only	✗	**0.476**	0.546	0.585	0.540	0.580	**0.580**	0.482	0.491	0.476	0.528 (±0.040)
	Score + Feedback	Low Quality (GPT-5.1)	0.445	0.560	0.603	0.579	0.611	0.550	0.561	0.568	0.542	0.558 (±0.032)
		High Quality (GPT-5.1)	0.438	0.592	0.601	0.585	0.606	0.545	0.548	0.575	0.555	0.561 (±0.038)
	Improvement (High Quality - Low Quality)		-1.57%	+5.71%	-0.33%	+1.04%	-0.82%	-0.91%	-2.32%	1.23%	+2.40%	0.52%
	Score + Feedback	Low Quality (FeedEval)	0.449	0.576	0.605	0.575	0.602	0.570	0.556	0.563	0.558	0.562 (±0.035)
		High Quality (FeedEval)	0.451	**0.601**	**0.612**	**0.587**	**0.617**	0.575	**0.598**	**0.598**	**0.579**	**0.580** (±0.037)
	Improvement (High Quality - Low Quality)		+0.45%	+4.34%	+1.16%	+2.09%	+2.49%	+0.88%	+7.55%	+6.22%	+3.76%	+3.22%
Qwen3-8B	Score Only	✗	**0.712**	0.693	0.696	0.677	0.710	**0.676**	0.674	0.681	0.684	0.689 (±0.022)
	Score + Feedback	Low Quality (GPT-5.1)	0.649	0.692	0.701	0.667	0.701	0.669	0.682	0.669	0.675	0.678 (±0.026)
		High Quality (GPT-5.1)	0.657	0.686	0.697	0.675	0.699	0.664	0.685	0.682	0.686	0.681 (±0.035)
	Improvement (High Quality - Low Quality)		+1.23%	-0.87%	-0.57%	1.20%	-0.29%	-0.75%	+0.44%	+1.94%	+1.63%	+0.43%
	Score + Feedback	Low Quality (FeedEval)	0.657	0.682	0.697	0.673	0.694	0.671	0.656	0.674	0.684	0.676 (±0.023)
		High Quality (FeedEval)	0.661	**0.699**	**0.709**	**0.683**	**0.719**	0.673	**0.694**	**0.688**	**0.698**	**0.692** (±0.020)
	Improvement (High Quality - Low Quality)		+0.61%	+2.49%	+1.72%	+1.49%	+3.60%	+0.30%	+5.79%	+2.08%	+2.05%	+1.58%

Table 5: Average essay scoring performance across all prompts for each trait on the ASAP++ dataset. Traits are predicted from right to left (←). We report five-fold averaged results with standard deviations (SD). The best performances are shown in **bold**, and the second-best are underlined for each LLM backbone.

GPT-5.1-selected high-quality feedback, further underscoring **FeedEval's stronger capability in identifying high-quality feedback**.

Compared to the score-only configuration, incorporating high-quality feedback slightly degrades performance on the Overall trait. This is because the dataset does not provide rubric score descriptions for this trait, leading us to replace its feedback with "NAN," which could hinder accurate scoring. For the Organization trait, the performance gap between high- and low-quality feedback is marginal, likely because its rubric descriptions are broadly defined compared to other traits, which limits the feedback's ability to capture structural features and leads to minimal semantic differences across feedback quality levels, as illustrated by the case study provided in Appendix J.4. Consequently, such **unclear feedback might have slightly degraded the scoring performance of Qwen3-8B on the Organization trait relative to the score-only setting**.

Overall, the results underscore the importance of well-defined, trait-specific feedback for accurate essay scoring and demonstrate the effectiveness of FeedEval for distinguishing essay feedback quality. Given its strong performance, we adopt Qwen3-8B as the backbone for subsequent experiments.[10]

5.2.2 Impact of FeedEval Dimension

Figure 3 summarizes the essay scoring performance of Qwen3-8B trained with high-quality feedback selected using different combinations of FeedEval dimensions. For one- and two-dimension settings, we average results across all possible single or pairwise combinations (see Appendix I for individual performance results). Overall, performance improves as more FeedEval dimensions are incorporated. **Using all three dimensions yields the best results across all traits**, while gains for the Overall and Organization traits remain marginal. This is likely due to the absence of feedback supervision for the Overall trait, as well as the difficulty of capturing structural features for the Organization trait, which is exacerbated by broad rubric descriptions with limited distinctions across score levels.

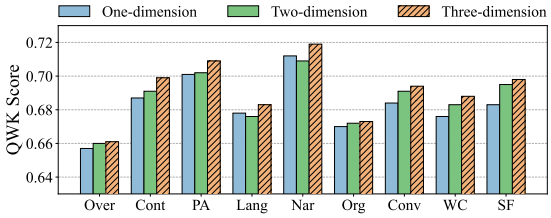

Figure 3: Average essay scoring performance across traits on ASAP++ for Qwen3-8B trained with high-quality feedback labels filtered by FeedEval using one, two, or all three dimensions.

5.3 Impact of Feedback Quality on Essay Revision (R3)

5.3.1 Essay Improvement after Revision

To evaluate essay revisions by small-sized LLMs, we use the Qwen3-8B essay scoring model trained with score-only labels (Avg. QWK = 0.689; Table 5). Figure 4 shows the average scores improved when Llama3-1B-Instruct and Qwen2-1.5B-Instruct revise essays using high- or low-quality

[10] Generating feedback before predicting scores consistently degraded performance compared to predicting scores first, as shown in the Appendix H.

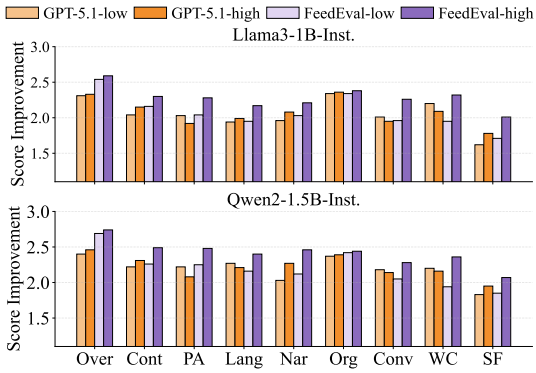

Figure 4: Average essay score improvement across traits on ASAP++ after revisions guided by feedback of high- and low-quality identified by FeedEval and GPT-5.1.

feedback identified by FeedEval and GPT-5.1.

From these results, we first observe that high-quality feedback filtered by FeedEval consistently yields larger revision gains, while improvements for the Organization trait remain marginal due to the lack of detailed rubric descriptions during feedback generation, as discussed in section 5.2. Furthermore, the revision gains from the high-quality feedback filtered by GPT-5.1 are not consistently larger than those from low-quality feedback, indicating limited capability of GPT-5.1 in distinguishing feedback quality. In addition, feedback identified as high-quality by FeedEval leads to greater revision gains than GPT-5.1-selected feedback. Overall, **these results demonstrate that FeedEval reliably identifies pedagogically high-quality feedback that enables more effective essay revisions**. Specific cases of essay feedback and revised essays are analyzed in Appendix J.5.

5.3.2 Human Evaluation of Feedback and Essay Reivison

Feedback	Quality of Feedback (1-5)			Quality of Revised Essay (%)
	D1	D2	D3	
High-quality	3.62	4.67	3.02	70.3
Low-quality	1.53	1.30	2.05	29.7

Table 6: Human evaluation comparing feedback quality and revised essay quality.

Table 6 reports human evaluations of both feedback quality and essay revision. Three educational experts evaluated high- or low-quality feedback filtered by FeedEval for 300 essays (100 per expert), rating feedback on a 5-point Likert scale across three pedagogically grounded feedback quality dimensions proposed by Steiss et al. (2024): faithfulness to essay (D1), usefulness for revision (D2), and rubric alignment (D3)[11]. The evaluation results show that the high-quality feedback consistently received higher scores compared to the low-quality feedback. The same experts also conducted pairwise comparisons of 300 essays revised by a small-sized LLM (Qwen2-1.5B-Instruct) using the two types of feedback, preferring essays revised with high-quality feedback significantly more often. **These results corroborate the automatic evaluation in section 5.3.1 and confirm that FeedEval effectively filters high-quality feedback, leading to improved downstream essay revision.**

6 Conclusion

In this paper, we introduce FeedEval, a novel LLM-based framework for evaluating the pedagogical quality of LLM-generated essay feedback along three dimensions: specificity, helpfulness, and validity. We validate FeedEval's alignment with human expert judgments and demonstrate how its evaluation scores can be used to filter high-quality feedback. Experiments on the ASAP++ dataset show that FeedEval closely matches expert evaluations and that models trained on FeedEval-filtered high-quality feedback achieve more accurate essay scoring than those trained on low-quality feedback. Moreover, essay revision experiments using small LLMs, together with human evaluations, confirm that the selected high-quality feedback is pedagogically more meaningful and effective. Finally, FeedEval consistently outperforms GPT-5.1 in distinguishing feedback quality, underscoring its effectiveness as a pedagogically aligned feedback evaluation framework and its potential to advance LLM-based automated essay assessment.

Limitations

In this study, we identify two primary limitations. First, the essay scoring performance of the LLMs used in our experiments may be influenced by the generation order of scores and feedback, reflecting the autoregressive nature of these models. Specifically, we observe that constructing labels to generate scores before feedback leads to better scoring performance than generating feedback first. This limitation, however, is not unique to our approach and is shared by many recent LLM-based models that jointly generate predictions and rationales.

[11] Details of each dimension are addressed in Appendix K

Second, our study focuses exclusively on English essay writing. To evaluate the generalizability of FeedEval to broader language education settings, future work should examine its applicability to essays written in other languages.

Ethical Statement

Our work used publicly available essay scoring benchmark datasets, including ASAP++, and did not pose any ethical concerns during experimentation. For human evaluation, we followed a well-established evaluation protocol in the literature, preventing possible ethical issues in the annotation process. Annotators were compensated at a rate approximately 30% higher than the average U.S. minimum wage.

Scientific Artifacts

The dataset (SpecEval) used to train the specificity evaluator was generated using GPT-4o via OpenAI's paid API services. See Appendix 3.2.1 for details.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korean government(MSIT) (No.RS-2022-NR068758).

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Barbara, Hamner Ben, Morgan Jaison, lynnvandev, and Shermis Mark. 2012. The hewlett foundation: Short answer scoring.

Shabnam Behzad, Omid Kashefi, and Swapna Somasundaran. 2024a. Assessing online writing feedback resources: Generative ai vs. good samaritans. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1638–1644.

Shabnam Behzad, Omid Kashefi, and Swapna Somasundaran. 2024b. Leaf: Language learners' english essays and feedback corpus. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 433–442.

Hamner Ben, Morgan Jaison, lynnvandev, Shermis Mark, and Vander Ark Tom. 2012. The hewlett foundation: Automated essay scoring.

Paul Black and Dylan Wiliam. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of personnel evaluation in education)*, 21(1):5–31.

Xiaoshu Chen, Sihang Zhou, Ke Liang, and Xinwang Liu. 2025. Distilling reasoning ability from large language models with adaptive thinking. *IEEE Transactions on Neural Networks and Learning Systems*.

Seong Yeub Chu, Jong Woo Kim, and Mun Yong Yi. 2025a. Think together and work better: Combining humans' and llms' think-aloud outcomes for effective text evaluation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

SeongYeub Chu, Jong Woo Kim, Bryan Wong, and Mun Yong Yi. 2025b. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5796–5814.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

David Dinucu-Jianu, Jakub Macina, Nico Daheim, Ido Hakimi, Iryna Gurevych, and Mrinmaya Sachan. 2025. From problem-solving to teaching problem-solving: Aligning LLMs with pedagogy using reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 272–292, Suzhou, China. Association for Computational Linguistics.

Heejin Do, Yunsu Kim, and Gary Lee. 2024. Autoregressive score generation for multi-trait essay scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1659–1666.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt-and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551.

Heejin Do, Sangwon Ryu, and Gary Geunbae Lee. 2025. Teach-to-reason with scoring: Self-explainable rationale-driven multi-trait essay scoring. *arXiv preprint arXiv:2502.20748*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Tak Yeon Lee, So-Yeon Ahn, and Alice Oh. 2024. RECIPE4U: Student-ChatGPT interaction dataset in EFL writing education. In *Proceedings of the*

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 13666–13676, Torino, Italia. ELRA and ICCL.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2023. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *Advances in Neural Information Processing Systems*, 36:48573–48602.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Prompting large language models for zero-shot essay scoring via multi-trait specialization. *arXiv preprint arXiv:2404.04941*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. Distilling chatgpt for explainable automated student answer assessment. *arXiv preprint arXiv:2305.12962*.

Shengjie Li and Vincent Ng. 2024. Automated essay scoring: A reflection on the state of the art. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17876–17888.

Shengjie Li and Vincent Ng. 2025. Graph-based multi-trait essay scoring. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33313–33339.

Zhenwen Liang, Dian Yu, Wenhao Yu, Wenlin Yao, Zhihan Zhang, Xiangliang Zhang, and Dong Yu. 2024. Mathchat: Benchmarking mathematical reasoning and instruction following in multi-turn interactions. *arXiv preprint arXiv:2405.19444*.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 204–221.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251.

Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.

Haile Misgna, Byung-Won On, Ingyu Lee, and Gyu Sang Choi. 2024. A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, 58(2):36.

Inderjeet Jayakumar Nair, Jiaye Tan, Xiaotian Su, Anne Gere, Xu Wang, and Lu Wang. 2024. Closing the loop: Learning to generate writing feedback via language model simulated student revisions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16636–16657.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506.

Alexander Scarlatos, Digory Smith, Simon Woodhead, and Andrew Lan. 2024. Improving the validity of automatically generated feedback via reinforcement learning. In *International Conference on Artificial Intelligence in Education*, pages 280–294. Springer.

Hyein Seo, Taewook Hwang, Yohan Lee, and Sangkeun Jung. 2025. FEAT: A preference feedback dataset through a cost-effective auto-generation and labeling framework for English AI tutoring. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 575–589, Vienna, Austria. Association for Computational Linguistics.

Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research*, 78(1):153–189.

Hwanjun Song, Taewon Yun, Yuho Lee, Jihwan Oh, Gihun Lee, Jason Cai, and Hang Su. 2025. Learning to summarize from llm-generated feedback. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 835–857.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*.

Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. 2024. Comparing the quality of human and chatgpt feedback of students' writing. *Learning and Instruction*, 91:101894.

Yongjie Wang, Chuan Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. *arXiv preprint arXiv:2205.03835*.

Zhaoyang Wang, Jinqi Jiang, Huichi Zhou, Wenhao Zheng, Xuchao Zhang, Chetan Bansal, and Huaxiu Yao. 2025. Verifiable format control for large language model generations. *arXiv preprint arXiv:2502.04498*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Prompt Templates

A.1 Prompt Design for Feedback Generation

We employ GPT-5.1 to generate essay feedback using multiple sources of information. For each essay, we provide the essay text, the associated prompt, and an excerpt when available. To introduce variation in the generated feedback, we additionally define three feedback-generation settings based on how human-assigned scores and rubric descriptions are incorporated.

A.1.1 Score+Rubric

In this setting, human-assigned scores and their corresponding rubric descriptions are explicitly included in the prompt. The prompt template is shown in Figure 5.

A.1.2 Score Only

This setting includes only the human-assigned score in the prompt, without providing the corresponding rubric description. The prompt template is shown in Figure 6.

A.1.3 Rubric Only

This setting includes rubric descriptions covering all score ranges for each trait, without providing human-assigned scores. The LLM evaluates the essay without knowing human-annotated scores. The prompt template is shown in Figure 7.

A.2 Prompt Design for Feedback Filtering by GPT-5.1

Figure 8 illustrates the prompt template used for filtering essay feedback quality via an LLM-as-a-judge framework, employing GPT-5.1 and Gemini-2.5-Pro. We leverage these LLMs through four distinct prompting strategies: zero-shot, zero-shot with CoT, few-shot, and few-shot with CoT. The colored segments are conditionally included based on the strategy: the black text serves as the base zero-shot prompt, while three author-provided selection examples are appended for the few-shot setting. The instruction "Think step by step" is included for CoT-based filtering. When filtering feedback based on a specific pedagogical dimension, the corresponding definition is inserted between the **[Condition]** and **(end of [Condition])** markers.

You are a member of the English essay writing test evaluation committee. Please, evaluate the given essay using following information.

[Prompt]
{prompt text}
(end of [Prompt])

[Excerpt]
{excerpt text}
(end of [Excerpt])

[Essay]
{essay text}
(end of [Essay])

[Scores]
Narrativity: 3
Language: 2
(. . .)
(end of [Scores])

[Rubric descriptions]
[Trait]
Narrativity
(end of [Trait])
The following is a rubric description in terms of the "Narrativity" trait.
Score 3: The response is interesting. Appropriate use of transition and . . .
[Trait]
Language
(. . .)
(end of [Rubric descriptions])

Refer to the provided **[Prompt]**, **[Excerpt]**, **[Scores]**, and **[Rubric descriptions]** to evaluate the given essay.
Your task is to analyze the reason why the essay got certain scores for each trait based on the analysis of the essay.

[Note]
I have made an effort to remove personally identifying information from the essays using the Named Entity Recognizer (NER). The relevant entities are identified in the text and then replaced with a string such as '@PERSON', '@ORGANIZATION', '@LOCATION', '@DATE', '@TIME', '@MONEY', '@PERCENT', '@CAPS' (any capitalized word) and '@NUM' (any digits). Please do not penalize the essay because of the anonymizations.
(end of [Note])

Q. Identify specific excerpts from the [Essay] that illustrate the strengths or weaknesses highlighted in the [Rubric descriptions] for each trait. Quote or summarize the relevant parts of the essay. Based on this analysis, rationalize the [Rubric descriptions] for each trait. If the [Rubric descriptions] for a given trait indicates that the writing is strong, provide only positive feedback. If it identifies weaknesses, provide a detailed analysis of the issue and suggest specific ways to improve it. Keep your response for each trait within three sentences, and do not include any specific scores in your analysis. Provide your answer in the following format:

{"trait 1": "evaluation for trait 1", "trait 2": "evaluation for trait 2", . . . }

Figure 5: Prompt template for feedback generation using both human-annotated **scores** and score descriptions of the **rubric**.

You are a member of the English essay writing test evaluation committee. Please, evaluate the given essay using following information.

[Prompt]
{prompt text}
(end of [Prompt])

[Excerpt]
{excerpt text}
(end of [Excerpt])

[Essay]
{essay text}
(end of [Essay])

Refer to the provided **[Prompt]** and **[Excerpt]** to evaluate the given essay. The following shows the scores of each trait provided by a human scorer.

[Scores]
Narrativity: 3
Language: 2
(. . .)
(end of [Scores])

Your task is to analyze the reason why the essay got certain scores for each trait.

[Note]
I have made an effort to remove personally identifying information from the essays using the Named Entity Recognizer (NER). The relevant entities are identified in the text and then replaced with a string such as '@PERSON', '@ORGANIZATION', '@LOCATION', '@DATE', '@TIME', '@MONEY', '@PERCENT', '@CAPS' (any capitalized word) and '@NUM' (any digits). Please do not penalize the essay because of the anonymizations.
(end of [Note])

Q. Identify specific excerpts from the [Essay] that illustrate the strengths or weaknesses for each trait. Quote or summarize the relevant parts of the essay. Based on your analysis, rationalize the score for each trait. If the writing is strong enough, provide only positive feedback. If there are some weaknesses, provide a detailed analysis of the issue and suggest specific ways to improve it. Keep your response for each trait within three sentences, and do not include any specific scores in your analysis. Provide your answer in the following format:

{"trait 1": "evaluation for trait 1", "trait 2": "evaluation for trait 2", . . . }

Figure 6: Prompt template for feedback generation using human-annotated **scores**.

You are a member of the English essay writing test evaluation committee. Please, evaluate the given essay using following information.

[Prompt]
{prompt text}
(end of [Prompt])

[Excerpt]
{excerpt text}
(end of [Excerpt])

[Rubric guidelines]
[Trait]
Narrativity
(end of [Trait])
[Trait Rubric]
Score 0: The response is irrelevant/incorrect/incomplete.
Score 1: The response is very uninteresting and disjointed and ...
(. . .)
(end of [Trait Rubric])

[Trait]
Language
(. . .)
(end of [Rubric guidelines])

Refer to the provided **[Prompt]** and **[Excerpt]** to evaluate the given essay.

[Essay]
{essay text}
(end of [Essay])

[Note]
I have made an effort to remove personally identifying information from the essays using the Named Entity Recognizer (NER). The relevant entities are identified in the text and then replaced with a string such as '@PERSON', '@ORGANIZATION', '@LOCATION', '@DATE', '@TIME', '@MONEY', '@PERCENT', '@CAPS' (any capitalized word) and '@NUM' (any digits). Please do not penalize the essay because of the anonymizations.
(end of [Note])

Q. Identify specific excerpts from the [Essay] that illustrate the strengths or weaknesses highlighted in the [Rubric guidelines] for each trait. Quote or summarize the relevant parts of the essay. Based on your analysis, rationalize your analysis for each trait. If the writing is strong enough, provide only positive feedback. If there are some weaknesses, provide a detailed analysis of the issue and suggest specific ways to improve it. Keep your response for each trait within three sentences, and do not include any specific scores in your analysis. Provide your answer in the following format:

{"trait 1": "evaluation for trait 1", "trait 2": "evaluation for trait 2", . . . }

Figure 7: Prompt template for feedback generation using the **rubric** guidelines.

On the other hand, when filtering feedback based on all three dimensions together, all the definitions are integrated into the template.

B Dataset Details

B.1 Datasets for Helpfulness Evaluator

The following three sources serve as the primary basis for building chosen–rejected pairs of the datasets for training the helpfulness evaluator:

- **RECIPE4U (Han et al., 2024):** A dataset collected from university-level English learners who iteratively revised their essays based on LLM-generated feedback. In our setting, feedback that students accepted and used for revision is labeled as chosen, whereas feedback that was not adopted is labeled as rejected.
- **FEAT (Seo et al., 2025):** A dataset in which students read English passages and wrote open-ended responses to comprehension questions. Human annotators ranked feedback—generated by both humans and LLMs—based on its accuracy and helpfulness in improving the student's answer, and we use these rankings to construct chosen–rejected pairs.
- **ASAP++ (Mathias and Bhattacharyya, 2018):** We use 45 human-written feedback instances from the ASAP++ dataset. Inspired by Behzad et al. (2024b), we prompt GPT-5.1 to generate revised versions of the human-written feedback, treating the revised feedback as the *chosen* example and the original human-written feedback as the *rejected* example.

B.2 Details of ASAP++ Dataset

Table 7 summarizes the ASAP++ dataset, including essay characteristics, evaluated traits, and score ranges for each prompt.

C Human Evaluation Details

We conducted various human evaluations with three teacher annotators holding master's degrees in English education.

C.1 Pairwise Comparison of Essay Feedback Quality across Specificity, Helpfulness, and Validity

Figure 9 illustrates the definitions and descriptive criteria for the three feedback evaluation dimensions. Prior to the main annotation, the three teacher annotators completed two training rounds with 10 practice pairs per round to establish a shared mental model of the evaluation criteria. Subsequently, we measured inter-rater reliability using Fleiss' Kappa on a subset of 30 pairs, observing agreement above 0.85, which indicates very high consistency. After verifying the consistency, the annotators labeled 150 pairs per dimension—*specificity*, *validity*, and *helpfulness*—resulting in a total of 450 annotated pairs across all dimensions.

C.2 Human-evaluation of Essay Feedback Filtered by FeedEval

Initially, the three teacher annotators completed two training sessions, each consisting of 10 practice essay feedback samples, to establish a shared understanding of the feedback evaluation criteria. Inter-rater reliability was then assessed using the intra-class correlation coefficient (ICC) on 30 essay feedback samples, indicating good agreement (D1=0.78, D2=0.72, D3=0.75).

C.3 Human-evaluation of Essay Revised by Small LLMs

The three teacher annotators were asked to select the better revised essay from each pair of revision outcomes. They again completed two training sessions with 10 essay pairs per session to calibrate their judgments, achieving high inter-rater consistency with a Fleiss' Kappa of 0.81 on 30 essay pairs.

D Implementation Details

Details of the configurations for training essay assessment LLMs and implementing essay revision LLMs are provided below:

D.1 FeedEval LLMs

We fine-tune Llama3-3B-Instruct[12], Qwen2-3B-Instruct[13], Phi-3-Mini-Instruct[14], and Gemma3-Instruct[15] models with DeepSpeed Stage-2 (Rasley et al., 2020) on eight NVIDIA A100 (80GB) GPUs.

[12] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[13] https://huggingface.co/Qwen/Qwen2.5-3B-Instruct

[14] https://huggingface.co/microsoft/Phi-3.5-mini-instruct

[15] https://huggingface.co/google/gemma-3-4b-it

You are an expert in filtering feedback data from multiple candidates based on specified conditions. Read the following conditions and select the feedback that best satisfies them.

[Condition]
1. The feedback {should/should not} quote parts of the essay that are relevant to evaluating the given traits (Specificity).
2. The feedback {should/should not} include actionable revision suggestions for improving the essay (Helpfulness).
3. The feedback {should/should not} align with the score descriptions in the rubric (Validity).
(end of [Condition])

[Examples]
Three examples of essays, scores, feedback candidates, and selected feedback
(end of [Examples])

[Essay]
{essay text}
(end of [Essay])

[Scores]
Narrativity: 3
Language: 2
(. . .)
(end of [Scores])

[Rubric descriptions]
[Trait]
Narrativity
(end of [Trait])
The following is a rubric description in terms of the "Narrativity" trait.
Score 3: The response is interesting. Appropriate use of transition and . . .
[Trait]
Language
(. . .)
(end of [Rubric descriptions])

[Feedback Candidates]

"feedback 1": "The essay is . . . "
"feedback 2": "The essay is . . . "
"feedback 3": "The essay is . . . "
"feedback 4": "The essay is . . . "
"feedback 5": "The essay is . . . "
"feedback 6": "The essay is . . . "
"feedback 7": "The essay is . . . "
"feedback 8": "The essay is . . . "

(end of [Feedback Candidates])

Think step by step.
Pick only one feedback without additional explanation:
ex) feedback 3

Figure 8: Prompt template for filtering essay feedback by using LLM-as-a-judge.

Prompt	# of Essays	Average Length	Essay Type	Grade Level	Traits	Score Range	
						Overall	Trait
P1	1,783	350	Argumentative	8	Over, Cont, WC, Org, SF, Conv	2 - 12	1 - 6
P2	1,800	350	Argumentative	10	Over, Cont, WC, Org, SF, Conv	1 - 6	1 - 6
P3	1,726	150	Source-Dependent	10	Over, Cont, PA, Nar, Lan	0 - 3	0 - 3
P4	1,772	150	Source-Dependent	10	Over, Cont, PA, Nar, Lan	0 - 3	0 - 3
P5	1,805	150	Source-Dependent	8	Over, Cont, PA, Nar, Lan	0 - 4	0 - 4
P6	1,800	150	Source-Dependent	10	Over, Cont, PA, Nar, Lan	0 - 4	0 - 4

Table 7: Statistics of the ASAP++ dataset. Traits include Overall (Over), Content (Cont), Word Choice (WC), Organization (Org), Sentence Fluency (SF), Conventions (Conv), Prompt Adherence (PA), Narrativity (Nar), and Language (Lang).

Specificity: Evaluates how concretely the feedback reflects the content of the essay.
- To what extent is the essay content explicitly referenced in the feedback?
- Is the essay content referenced evenly across the sentences of the feedback?
- Does the feedback refer to multiple parts of the essay in a balanced manner?

Helpfulness: Evaluates the extent to which the feedback supports improvement of the learner's essay.
- Does the feedback identify aspects of the essay that need improvement?
- Does the feedback provide sufficient information to help the student improve the essay?

Validity: Evaluates how well the feedback reflects the level of the essay score based on the rubric criteria.
- To what extent does the feedback reflect the rubric criteria corresponding to the essay score?
- How accurately does the feedback use expressions from the rubric criteria corresponding to the essay score?

Figure 9: Definitions and detailed descriptions of feedback evaluation dimensions.

Training is conducted for 5 epochs (100 steps) using AdamW with a batch size of 4, an initial learning rate of 1e-5. The margin hyperparameter for training the evaluator of specificity and helpfulness is set to 0.5. All other hyperparameters follow default settings.

D.2 Essay Assessment LLMs

We fine-tune Llama3-8B-Instruct[16] and Qwen3-8B[17] models using LoRA (Hu et al., 2022) with DeepSpeed Stage-2 (Rasley et al., 2020) on eight NVIDIA A100 (80GB) GPUs. Training is conducted for 5 epochs (100 steps) using AdamW with a batch size of 4, an initial learning rate of 1e-4, a weight decay of 0.05, and early stopping with a patience of 2. All other hyperparameters follow default settings.

D.3 Essay Revision LLMs

For essay revision, we prompt Llama3-1B-Instruct[18] and Qwen2-1.5B-Instruct[19] models to revise essays based on provided feedback. We set the temperature to 0.7 and generate up to 1,000 new tokens.

E Specificity Evaluator Trained on Llama-generated SpecEval Dataset

We generate a new SpecEval dataset using Llama3-70B [20]. Then we train Llama3-3B-Instruct on the SpecEval and compare its agreement with expert annotations as well as with the Llama3-3B-Instruct trained on GPT-4o-generated SpecEval. As shown in Table 8, the evaluators trained on the

[16] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[17] https://huggingface.co/Qwen/Qwen3-8B

[18] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

[19] https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct

[20] https://huggingface.co/meta-llama/Meta-Llama-3-70B

LLaMA-synthesized data demonstrated comparable human–LLM agreement in terms of accuracy and F1 score to its GPT-4o-based counterpart. Furthermore, both evaluators achieved high agreement scores exceeding 0.8 in both metrics.

Model	GPT vs. Human		Llama vs. Human		GPT vs. Llama	
	Acc.	F1	Acc.	F1	Acc.	F1
Llama3-3B-Inst.	0.820	0.880	0.800	0.864	0.824	0.884
Qwen2-3B-Inst.	0.807	0.870	0.778	0.858	0.812	0.879
Phi-3-Mini	0.811	0.860	0.793	0.852	0.826	0.863
Gemma3-Inst.	**0.832**	**0.893**	**0.827**	**0.881**	**0.840**	**0.902**

Table 8: Agreement of specificity evaluators trained on SpecEval datasets synthesized by Llama3-70B (Llama) or GPT-4o (GPT), compared with human experts and between the two evaluators. All 3B-scale models are fine-tuned. The best results are in bold, and the second-best are underlined.

F Analysis of Potential Conflicts Between Evaluation Dimensions

Dimension	Specificity	Helpfulness	Validity
Specificity	-	0.597	0.535
Helpfulness	0.597	-	0.511
Validity	0.535	0.511	-

Table 9: Pearson Correlation Coefficients Among Evaluation Dimensions

To evaluate potential conflicts among the three evaluation dimensions—*specificity*, *helpfulness*, and *validity*—the Pearson correlation coefficients between the scores generated by each dimension-specific evaluation model (Llama3-3B-Instruct) were analyzed. As presented in Table 9, the results indicate moderate positive correlations across all pairs ($0.40 < r < 0.60$). This suggests that the dimension scores exhibit some linear association and are not in opposing directions, while still capturing distinguishable aspects.

G Impact of FeedEval on ASAP-SAS Dataset

G.1 ASAP-SAS Dataset

To examine generalizability, we additionally utilize the ASAP-SAS dataset (Barbara et al., 2012)[21], which contains open-ended student responses across multiple subjects. Since no publicly available dataset other than the ASAP++ provides multi-trait essay scores directly linked to scoring rubrics, we adopt ASAP-SAS as a practical alternative, despite it offering only overall quality labels. We select five prompts related to essay writing, and use this dataset as a complementary benchmark to evaluate the generalizability of FeedEval in essay assessment settings.

Dataset	Prompt	# Essays	Score Range
ASAP-SAS	3	2214	0 - 3
	4	1952	
	7	2398	
	8	2398	
	9	2397	

Table 10: Statistics of the ASAP-SAS dataset.

G.2 Essay Scoring Performance

Table 11 reports essay scoring performance of models trained on the ASAP-SAS dataset to jointly generate scores and feedback. Consistent with ASAP++, models trained on high-quality feedback filtered by FeedEval outperform those trained on low-quality feedback across all prompts. In contrast, feedback labeled as high-quality by GPT-5.1 does not consistently yield better performance than its low-quality counterparts. Moreover, models trained with FeedEval-selected high-quality feedback consistently surpass those trained with GPT-5.1-selected feedback. These results demonstrate that FeedEval's feedback quality assessment generalizes beyond ASAP++ to other essay datasets.

G.3 Essay Improvement after Revision

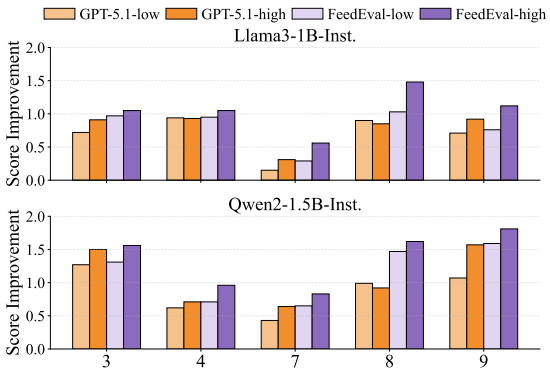

Figure 10: Average essay score improvement across prompts on ASAP-SAS after revisions guided by feedback of high- and low-quality identified by FeedEval and GPT-5.1.

Figure 10 shows average prompt-level score improvements when Llama3-1B-Instruct and Qwen2-

[21] https://www.kaggle.com/competitions/asap-sas

LLM	Assessment	Feedback Quality	Prompt 3	4	7	8	9	Avg ↑ (SD ↓)
Llama3-8B-Inst.	Score Only	✗	0.556	0.551	0.534	0.527	0.634	0.560 (±0.037)
	Score + Feedback	Low Quality (GPT-5.1)	0.589	0.602	0.498	0.551	0.695	0.587 (±0.027)
		High Quality (GPT-5.1)	0.641	0.597	0.539	0.549	0.709	0.607 (±0.029)
	Improvement (High Quality - Low Quality)		+8.83%	-0.83%	+8.23%	-0.36%	+2.01%	+3.41%
	Score + Feedback	Low Quality (FeedEval)	0.617	0.604	0.521	0.563	0.704	0.602 (±0.030)
		High Quality (FeedEval)	**0.661**	**0.610**	**0.551**	**0.595**	**0.729**	**0.629** (±0.035)
	Improvement (High Quality - Low Quality)		+7.13%	+0.99%	+5.76%	+5.68%	+3.55%	+4.55%
Qwen3-8B	Score Only	✗	0.570	0.585	0.483	0.542	0.691	0.574 (±0.027)
	Score + Feedback	Low Quality (GPT-5.1)	0.565	0.581	0.454	0.537	0.681	0.564 (±0.031)
		High Quality (GPT-5.1)	0.580	0.577	0.479	0.549	0.673	0.572 (±0.037)
	Improvement (High Quality - Low Quality)		+2.65%	-0.69%	+5.51%	+2.23%	-1.17%	+1.42%
	Score + Feedback	Low Quality (FeedEval)	0.582	0.577	0.491	0.553	0.706	0.582 (±0.035)
		High Quality (FeedEval)	**0.606**	**0.617**	**0.521**	**0.563**	**0.726**	**0.607** (±0.030)
	Improvement (High Quality - Low Quality)		+4.12%	+6.93%	+6.11%	+1.81%	+2.83%	+4.26%

Table 11: Average essay scoring performance for each prompt on the ASAP-SAS dataset. We report five-fold averaged results with standard deviations (SD). The best performances are shown in **bold**, and the second-best are underlined for each LLM backbone.

1.5B-Instruct revise essays using high- or low-quality feedback identified by FeedEval and GPT-5.1. Feedback filtered as high-quality by FeedEval consistently yields larger revision gains across traits, whereas GPT-5.1-filtered feedback does not show consistent improvements over its low-quality counterpart, indicating limited discriminative ability. Moreover, FeedEval-selected high-quality feedback leads to greater gains than GPT-5.1–selected feedback. Overall, these results demonstrate that FeedEval reliably identifies high-quality feedback that enables more effective essay revisions on ASAP-SAS, supporting its generalizability.

H Essay Assessment Performance with Reversely Constructed Labels

Table 12 reports the essay scoring performance of Llama3-8B-Instruct and Qwen3-8B trained on high- and low-quality feedback datasets filtered by FeedEval, where models are trained to generate feedback before predicting scores. While training on high-quality feedback still yields consistent improvements over low-quality feedback across traits, overall scoring performance is lower than that of models trained to predict scores first, indicating that the order of score and feedback generation affects assessment accuracy (Do et al., 2025). This suggests that LLMs benefit more from post-thinking mechanisms than from pre-thinking mechanisms in essay scoring tasks (Chen et al., 2025).

I Essay Scoring Performance under Different FeedEval Dimension Configurations

Table 13 reports the essay scoring performance of Qwen3-8B trained on high-quality feedback filtered by FeedEval under different dimension configurations. For multi-dimensional configuration, feedback is selected based on the average FeedEval score computed over the corresponding subset of dimensions.

J Additional Experiments

J.1 Essay Scoring Performance averaged across the traits for each prompt

Table 14 reports the essay scoring performance of Llama3-8B-Instruct and Qwen3-8B trained on different feedback quality, averaged across traits for each prompt. Models trained with high-quality feedback filtered by FeedEval consistently outperform those trained with low-quality feedback across all prompts, demonstrating the effectiveness of high-quality feedback as supervision for essay scoring. In contrast, models trained on feedback filtered as high-quality by GPT-5.1 do not yield consistent performance gains over their low-quality counterparts. Moreover, models trained on FeedEval-selected high-quality feedback consistently outperform those trained on GPT-5.1-selected high-quality feedback, highlighting FeedEval's superior ability to assess and filter pedagogi-

LLM	Assessment	Feedback Quality	Traits (Prediction Order: ←) Over	Cont	PA	Lang	Nar	Org	Conv	WC	SF	Avg↑ (SD↓)
Llama3-8B-Inst.	Score+Feedback	Low Quality	0.455	0.504	0.561	0.534	0.579	0.453	0.443	0.496	0.502	0.503 (±0.042)
		High Quality	0.464	0.508	0.590	0.538	0.600	0.470	0.461	0.530	0.527	0.521 (±0.032)
	Improvement (High Quality - Low Quality)		+1.98%	+0.79%	+5.17%	+0.75%	+3.63%	+3.75%	+4.06%	+6.85%	+4.98%	+3.56%
Qwen3-8B	Score+Feedback	Low Quality	0.637	0.607	0.613	0.614	0.624	0.664	0.691	0.646	0.684	0.642 (±0.024)
		High Quality	0.648	0.626	0.618	0.662	0.626	0.679	0.695	0.650	0.698	0.656 (±0.026)
	Improvement (High Quality - Low Quality)		+1.73%	+3.13%	+0.82%	+7.82%	+0.32%	+2.26%	+0.58%	+0.62%	+2.05%	+2.11%

Table 12: Average essay scoring performance across all prompts for each trait on the ASAP++ dataset. Traits are predicted from right to left (←). We report five-fold averaged results with standard deviations (SD).

# of Dimensions	Dimension	Traits (Prediction Order: ←) Over	Cont	PA	Lang	Nar	Org	Conv	WC	SF	Avg↑ (SD↓)
One	Specificity	0.657	0.689	0.700	0.678	0.712	0.669	0.688	0.672	0.681	0.683 (±0.021)
	Helpfulness	0.652	0.686	0.701	**0.683**	0.717	0.668	0.681	0.671	0.685	0.683 (±0.024)
	Validity	0.661	0.685	0.703	0.672	0.707	0.673	0.683	0.684	0.682	0.683 (±0.025)
Two	Specificity+Helpfulness	0.654	0.691	0.698	0.673	0.709	**0.675**	0.686	0.681	0.685	0.684 (±0.020)
	Specificity+Validity	0.662	0.692	0.701	0.678	0.708	0.672	0.691	0.683	0.696	0.687 (±0.025)
	Helpfulness+Validity	**0.664**	0.69	0.706	0.677	0.71	0.67	**0.697**	0.685	**0.703**	0.689 (±0.029)
Three	Specificity+Helpfulness+Validity	0.661	**0.699**	**0.709**	**0.683**	**0.719**	0.673	0.694	**0.688**	0.698	**0.692** (±0.020)

Table 13: Average essay scoring performance on the ASAP++ dataset using high-quality feedback filtered by different FeedEval dimension configurations, averaged across all prompts for each trait. Traits are predicted from right to left (←). We report five-fold averaged results with standard deviations (SD). The best performances are shown in **bold**, and the second-best are underlined.

cally useful feedback.

J.2 Alignment of FeedEval Across Model Scales

To further examine FeedEval's alignment with human experts in judging essay feedback quality, we compare Llama3-Instruct models of different parameter sizes fine-tuned on dimension-specific datasets. As shown in Table 15, alignment with human experts improves as model size increases. Ultimately, we adopt the Llama3-3B-Instruct model as the backbone for FeedEval, as it offers a reasonable balance between strong alignment with human experts and computational efficiency.

J.3 Error Analysis

We analyze error cases in which essay scoring LLMs fail to generate outputs in the expected score-feedback JSON format when trained with feedback of varying quality. Such errors are commonly observed when LLMs are required to generate text in a predefined structured format (Wang et al., 2025; Do et al., 2024). Aggregating errors across all folds, we observe that on the ASAP++ dataset, Llama3-8B-Instruct exhibits a total error rate of 1.6% across five folds, whereas Qwen3-8B shows a substantially lower error rate of 0.3%. In contrast, no formatting errors are observed for any model on the ASAP-SAS dataset. Since the primary focus of this study is essay scoring performance rather than error-rate reduction, we exclude these error cases when computing QWK scores.

J.4 Case Study of FeedEval-filtered Feedback

Table 16 compares high- and low-quality feedback from three FeedEval dimensions. Text highlighted in yellow indicates excerpts that directly reference specific parts of the essay (specificity). Text highlighted in blue represents actionable revision suggestions (helpfulness). Text highlighted in green marks content aligned with the rubric description (validity). Overall, the high-quality feedback covers a broader range of elements related to specificity, helpfulness, and validity than the low-quality feedback.

J.5 Case Study of Revised Essays

Table 17 presents revised essays produced by Qwen2-1.5B-Instruct based on feedback of different quality for the same original human-written essay, along with the corresponding scores assigned by an automated scoring model (Qwen3-8B). In the essays, revisions corresponding to each trait are highlighted using the same color assigned to that trait. For the Organization trait, no substantial differences are observed between high- and low-quality feedback, and the resulting revised essays receive identical scores for this trait.

LLM	Assessment	Feedback Quality	Prompt 1	2	3	4	5	6	Avg
Llama3-8B-Inst.	Score Only	✗	0.430	0.519	**0.600**	0.659	**0.567**	0.514	0.548 (0.042)
	Score + Feedback	Low Quality (GPT-5.1)	0.455	0.526	0.555	0.675	0.531	0.585	0.555 (0.041)
		High Quality (GPT-5.1)	0.443	0.517	0.567	0.668	0.544	0.584	0.554 (0.043)
	Improvement (High Quality - Low Quality)		-2.64%	-1.71%	+2.16%	-1.04%	+2.45%	-0.17%	-0.12%
	Score + Feedback	Low Quality (FeedEval)	0.479	0.537	0.553	0.674	0.540	0.564	0.558 (0.047)
		High Quality (FeedEval)	**0.492**	**0.544**	0.588	**0.689**	0.547	**0.587**	**0.575** (0.040)
	Improvement (High Quality - Low Quality)		+2.71%	+1.30%	+6.33%	+2.23%	+1.30%	+4.08%	+2.99%
Qwen3-8B	Score Only	✗	**0.687**	0.662	0.695	0.747	0.671	0.685	0.690 (0.025)
	Score + Feedback	Low Quality (GPT-5.1)	0.618	0.648	0.695	0.756	0.663	0.702	0.680 (0.029)
		High Quality (GPT-5.1)	0.631	0.645	0.701	0.753	0.671	0.689	0.682 (0.034)
	Improvement (High Quality - Low Quality)		+2.10%	-0.46%	+0.86%	-0.40%	+1.21%	-1.85%	+0.20%
	Score + Feedback	Low Quality (FeedEval)	0.638	0.647	0.701	0.752	0.667	0.687	0.682 (0.027)
		High Quality (FeedEval)	0.639	**0.665**	**0.705**	**0.765**	**0.675**	**0.711**	**0.693** (0.021)
	Improvement (High Quality - Low Quality)		+0.16%	+2.78%	+0.57%	+1.73%	+1.20%	+3.49%	+1.66%

Table 14: Average essay scoring performance across all traits for each prompt on the ASAP++ dataset. We report five-fold averaged results with standard deviations (SD). The best performances are shown in **bold**, and the second-best are underlined for each LLM backbone.

Model	Specificity		Helpfulness		Validity	
	Acc.	F1	Acc.	F1	Acc.	F1
Llama3-Inst. (1B)	0.793	0.857	0.751	0.838	0.682	0.564
Llama3-Inst. (3B)	0.820	0.880	0.864	0.912	0.835	0.709
Llama3-Inst. (8B)	**0.829**	**0.897**	**0.881**	**0.926**	**0.858**	**0.722**

Table 15: Human alignment of FeedEval using Llama3-Instruct across different parameter scales (pairwise Acc./F1). The best performances are shown in **bold**, and the second-best are underlined.

In contrast, for the remaining four traits, differences in feedback quality are clearly reflected in the revised essays and are further manifested in score differences. First, for the Sentence Fluency trait, both high- and low-quality feedback refer to the same portions of the original essay when explaining the issues, and the opening sections of the revised essays produced under both feedback conditions are very similar. However, although both feedback types identify areas for improvement, the high-quality feedback offers more concrete and actionable guidance than the low-quality feedback. This difference is reflected in the revisions: while not all revised segments can be explicitly highlighted, the essay revised using high-quality feedback exhibits clearer improvements in the Sentence Fluency trait, including more varied sentence structures, and consequently receives a higher score for the corresponding trait. In contrast, the essay revised using low-quality feedback contains repeated expressions that convey meanings similar to those in preceding sentences (■).

For the Word Choice trait, issues explicitly identified in each feedback are appropriately revised in the corresponding essays. However, aspects mentioned in the high-quality feedback but omitted in the low-quality feedback remain unaddressed in the revised essay guided by low-quality feedback, as shown in (⋆).

Next, for the Conventions trait, the high-quality feedback explicitly references specific contents of the essay, whereas the low-quality feedback lacks such specificity. As a result, the essay revised using high-quality feedback exhibits more varied sentence structures by rephrasing parts of the initial essay that contained grammatical errors and punctuation mistakes. In contrast, although revisions based on low-quality feedback reduce some conventions-related errors, they primarily involve deleting problematic sentences without introducing newly reformulated content, resulting in relatively limited sentence variation. Not all revised segments can be explicitly highlighted.

Finally, for the Content trait, the high-quality feedback provides more concrete and actionable guidance on how to improve the essay than the low-quality feedback. Consequently, the essay revised with high-quality feedback presents more explicit and well-supported evidence to substantiate the author's main arguments.

Trait	Sentence Fluency (Score: 3/6), Word Choice (Score: 3/6), Conventions (Score: 3/6), Organization (Score: 4/6), Content (Score: 3/6)
Essay	In the @LOCATION1 we have the technology of a computer. Some say that the computers are good for the society. I disagree, I believe that it is bad for a few reasons. Some of the reasons are obesity, cramps, more sexual harrassment and even cyber bullying. First if people don't get off the computers and go out to exercise then it will cause obesity. Think of it this way, if you watch a kid that didn't have to go to school his entire life and he started off at @NUM1 pounds. The only thing the kid will want to do is play on the computer and he will gain weight. Next, tip-top-tip-top, that's all you hear when a kid is on the computer, @CAPS1 teens come home from school and go straight to the computer and don't get off about on a school night. These are the kids that are @CAPS1 likely to get bad cramps. they get the cramps form typing to fast, hard and too long. They also can get the cramps from just sitting down for to long then trying to get up but can't because it hurts to much to decide to move anywhere. There's also a lot more of sexual harassment that is going on some kids goin to a chat room to talk to their friends about some great news, but when they get out of the chat room they are all mad and pissed off. The reason is because someone was making fun of them and they didn't like it. This happens maybe close to @PERCENT1 of the time that teens and even parents go into chat rooms online. This is the third reason on why computers have an huge effect on people. The last reason that computers have a huge effect on people is because of cyberbullying. cyberbullying is when someone bullies you but just online instead. Have you realized that children became poor or sick because they dont want to be hurt or killed over something very stupid? Well, I have and it's not nice, and now the god thing is if a kid goes to school online he/she @MONTH1 get cyberbullied because he's not smart enough or if he/she don't have the money to go to school. The cyberbullying normally occures in a chat room online. There is where all of the family will either become broke with no money or food. They @MONTH1 even lose their kid or the house that they live in because someone wanted to act all big and bad to be popular in school. In conclusion, these are only a few ways that computers effects people. The obesity, cramps, sexual harrassment and cyberbullying are not a joke and it's not fun to get done to you. Now it's your turn to tell us what you think about the computers. Do you think computers are good and benefits society or is it bad a rot little kids brains to make their lives horrible for them to live with?
High-quality Feedback	**Sentence Fluency**: The essay contains sentences such as 'First if people don't get off the computers and go out to exercise then it will cause obesity,' which lacks fluidity and clearer punctuation. Many sentences feature repetitive patterns and awkward constructions like 'The obesity, cramps, sexual harassment and cyberbullying are not a joke and it's not fun to get done to you.' Improvement can be made by varying sentence structures and adding transitional elements for smoother flow. **Word Choice**: Word choices such as 'it's not nice' and 'act all big and bad' lack specificity and precision, making the language seem mundane. Expressions like 'very stupid' and 'rot little kids brains' do not effectively convey the intended seriousness of the issues. Enhancing the vocabulary with more precise and varied language would make the argument more compelling. **Conventions**: The essay exhibits issues with conventions, notably in the lack of proper capitalization, punctuation, and spelling errors, such as 'sexual harrassment,' 'occures,' and 'dont.' There are frequent grammatical errors that disrupt readability and comprehension. To enhance conventions, careful proofreading and editing should focus on correcting spelling, capitalization, punctuation, and grammatical errors. **Organization**: The essay has a basic organizational structure with an introduction, body, and conclusion but does not fully function as a cohesive letter to the editor. Ideas are somewhat scattered throughout the paragraphs, affecting the clear flow of arguments. Refining the format to more clearly differentiate and develop each argument in separate, well-connected paragraphs and ensuring it adheres to the letter format will enhance organization. **Content**: The essay attempts to cover main ideas such as obesity and cyberbullying but does so in a simplistic and unstructured manner. Details provided, like 'someone wanted to act all big and bad to be popular in school,' lack depth and sometimes drift off-topic. Expanding on each issue with more thoughtful and specific examples would provide a deeper exploration of the problems discussed.
Low-quality Feedback	**Sentence Fluency**: The sentence fluency score reflects a mix of clear and awkward sentences throughout the essay such as 'The obesity, cramps, sexual harassment and cyberbullying are not a joke and it's not fun to get done to you.' While the writer attempts to use a conversational style, many sentences are fragmented or run-on, making it difficult to follow the argument smoothly. The lack of variety in sentence structure further limits the fluid reading experience. **Word Choice**: The word choice in the essay is basic and repetitive, which does not effectively convey a sophisticated understanding of the topic. There are attempts to use descriptive language, but the expression like 'rot little kids brains' is often imprecise or informal, detracting from the persuasive impact. Additionally, some words and phrases are used inaccurately, weakening the overall argument. **Conventions**: The essay contains several grammatical errors, such as incorrect verb forms, punctuation mistakes, and capitalization issues, which disrupt the readability. These errors suggest a lack of careful proofreading and editing. While some sentences are free of errors, the frequency of mistakes throughout the essay impacts the clarity and professionalism of the writing. To enhance conventions, the writer should focus on correcting spelling, capitalization, and grammatical errors. **Organization**: The organization of the essay shows a logical progression of ideas with a clear introduction, body, and conclusion. The writer presents several reasons supporting their opinion against the use of computers, although transitions between ideas could be smoother. Despite some structural weaknesses, the overall organization helps convey the writer's main points effectively. **Content**: The content of the essay includes several arguments against the use of computers, such as obesity and cyberbullying, but lacks depth and supporting evidence. The points are presented without substantial elaboration or examples that would strengthen the argument. Consequently, the essay provides a basic exploration of the topic without fully persuading the reader.
Rubric Description	**Sentence Fluency (Score 3)**: The writing tends to be mechanical rather than fluid. Occasional awkward constructions may force the reader to slow down or reread. The writing is characterized by -some passages that invite fluid oral reading; however, others do not. -some variety in sentence structure, length, and beginnings, although the writer falls into repetitive sentence patterns. **Word Choice (Score 3)**: Language lacks precision and variety, or may be inappropriate to audience and purpose in places. The writer does not employ a variety of words, producing a sort of 'generic' paper filled with familiar words and phrases. The writing is characterized by -words that work, but that rarely capture the reader's interest. -expression that seems mundane and general; slang, if used, does not seem purposeful and is not effective. -attempts at colorful language that seem overdone or forced. **Conventions (Score 3)**: The writing demonstrates limited control of standard writing conventions (e.g., punctuation, spelling, capitalization, grammar and usage). Errors begin to impede readability. The writing is characterized by -some control over basic conventions; the text may be too simple or too short to reveal mastery. -end-of-sentence punctuation that is usually correct; however, internal punctuation contains frequent errors. -spelling errors that distract the reader; misspelling of common words occurs. -errors in grammar and usage that do not block meaning but do distract the reader. **Organization (Score 4)**: The essay shows satisfactory organization. It contains a basic introduction, body and conclusion. **Content (Score 3)**: The reader can understand the main ideas, although they may be overly broad or simplistic, and the results may not be effective. Supporting detail is often limited, insubstantial, overly general, or occasionally slightly off-topic. The writing is characterized by -an easily identifiable purpose and main idea(s). -predictable or overly-obvious main ideas; or points that echo observations heard elsewhere; or a close retelling of another work. -support that is attempted, but developmental details are often limited, uneven, somewhat off-topic, predictable, or too general (e.g., a list of underdeveloped points).

Table 16: Case study illustrating an essay, high- and low-quality feedback for the Sentence Fluency, Word Choice, Conventions, Organization, and Content traits filtered by FeedEval, and the rubric descriptions corresponding to human-annotated scores. Due to space constraints, we present only the portions of the rubric descriptions that align with the feedback.

Initial Trait Scores	Sentence Fluency (Score: 1/6), Word Choice (Score: 1/6), Conventions (Score: 1/6), Organization (Score: 1/6), Content (Score: 1/6)
Topic Prompt	More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and even allow people to talk online with other people. Others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends. Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.
Initial Human-written Essay	More and more people use computers, they love computers for place that are far or the news. Computers are good for everyone in this world. Know one can live with out computers. There's others dat don't like computer for nothing. I know that there's family or friends that be having computers they write each others like in myspace u could have friend from everywhere or family's. But computers are spending to many times on their computers.
High-quality Feedback	**Sentence Fluency**: The essay lacks fluidity and coherence, with sentences such as 'Computers are good for everyone in this world. Know one can live with out computers.' that are choppy and disconnected. The structure is awkward, making it difficult to follow the argument. To improve, the writer should work on creating smooth transitions between ideas and constructing more complex and varied sentences. **Word Choice**: The essay's word choice is limited and imprecise, with phrases like 'Computers are good for everyone' lacking depth and specificity. Misused words such as 'Know one' instead of 'No one' further obscure clarity. Improvement can be achieved by using precise vocabulary and ensuring correct word usage to convey ideas more effectively. **Conventions**: Numerous grammatical errors such as 'Know one can live with out computers.' and 'There's others dat don't like computer' impair readability. The frequent spelling errors ('dat') and punctuation issues indicate a lack of control over conventions. Thorough proofreading and editing to correct these errors would improve readability. **Organization**: The essay lacks a clear organizational structure, evidenced by the mix of different ideas in sentences like 'There's others dat don't like computer for nothing.' The lack of clear paragraphs and transitions leads to a disjointed reading experience. Establishing a clear introduction, body, and conclusion and organizing related ideas into paragraphs would improve coherence. **Content**: The content is underdeveloped with vague ideas, lacking depth and specificity in addressing the essay prompt. Statements like 'Computers are good for everyone in this world' are too broad and unsupported by examples or discussion. To enhance content, focus on presenting a clear opinion with well-developed arguments supported by concrete details and examples.
Revised Essay with High-quality Feedback	More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and even allow people to talk online with other people. Others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends. However, I strongly disagree with this viewpoint. I firmly believe that computers can be a powerful tool that has a profound impact on people's lives. Computers have numerous benefits. They have made it possible for people to access vast amounts of information, allowing them to learn about various subjects and cultures. The internet has also enabled people to connect with others across the globe, fostering global understanding and collaboration. Furthermore, computers have greatly improved communication skills, enabling people to express themselves more effectively and engage in meaningful conversations. In addition, computers have the power to enhance creativity and productivity. They offer a wide range of tools and software that can be used for creative pursuits, such as writing, art, and music. Moreover, computers have improved access to healthcare, education, and employment opportunities, which have greatly benefited society as a whole. I firmly believe that computers are a valuable resource that can bring people together and improve their lives in numerous ways. While there are valid concerns about the negative effects of excessive computer use, I firmly disagree that this outweighs the benefits. By harnessing the power of technology, we can create a brighter, more connected, and more prosperous world for everyone. **(Sentence Fluency: 4/6, Word Choice: 4/6, Conventions: 4/6, Organization: 4/6, Content: 4/6)**
Low-quality Feedback	**Sentence Fluency**: The essay is difficult to follow, with awkward constructions such as 'Computers are spending to many times on their computers.' The sentences lack flow and structure, making the text jarring and irregular. To improve, the writer should focus on enhancing the fluidity of sentence structures. **Word Choice**: The essay employs vague and imprecise language, lacking specificity and reducing clarity for the readers. The limited vocabulary limits clarity and weakens the impact of the essay. For improvement, the writer should use precise and varied vocabulary to better convey the intended message. **Conventions**: There are numerous grammatical errors and incorrect convention usage, as seen in the inconsistency with punctuation and capitalization. These errors distract from the readability. Focus on proofreading to correct the grammatical mistakes and incorrect punctuation to improve clarity. **Organization**: The essay lacks a clear organizational structure, with ideas not clearly separated or presented in a logical sequence. It is written as a single fragmented paragraph without identifying distinct ideas or paragraph boundaries. The absence of clear introduction, body, and conclusion and overall coherence hinders the reader's ability to follow the argument. **Content**: The essay presents a limited and underdeveloped exploration of the topic, with vague and general statements such as 'Computers are good for everyone in this world.' The main ideas are not clearly articulated and underdeveloped, limiting the essay's persuasiveness.
Revised Essay with Low-quality Feedback	More and more people use computers, (⋆) and computers are good for everyone in this world. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and allow people to talk online with other people. However, others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends. (■) The widespread use of computers has led to a loss of face-to-face interactions and social skills, which are essential for building strong relationships and understanding each other. While computers can be a valuable tool for learning and communication, they can also have negative consequences. The constant exposure to screens and the lack of physical activity can lead to a sedentary lifestyle, which can have serious health risks. Furthermore, excessive computer use can cause social isolation and decreased social skills, as people spend more time interacting with computers than with others. Despite these concerns, the benefits of technology extend beyond just the individual. Computers have the power to bring people together and facilitate global communication, allowing people to connect with others across the world. They have also enabled people to access a vast amount of information and knowledge, which can be a valuable resource for learning and personal growth. **(Sentence Fluency: 3/6, Word Choice: 3/6, Conventions: 3/6, Organization: 4/6, Content: 3/6)**

Table 17: Case study illustrating essays revised using high- and low-quality feedback filtered by FeedEval across the Sentence Fluency, Word Choice, Conventions, Organization, and Content traits, along with scores assigned by an automated essay scoring model.

J.6 Essay Improvement after Revision using Feedback Provided by Trained Models

To evaluate the pedagogical effectiveness of essay feedback generated by models trained on high- and low-quality feedback filtered by FeedEval and GPT-5.1, we conduct essay revision experiments using small-sized LLMs (Llama3-1B-Instruct and Qwen2-1.5B-Instruct) together with a fine-tuned essay scoring model (Qwen3-8B). All experimental settings are identical to those in section 5.3.1, except for the source of feedback used for revision.

Specifically, in this experiment, the small LLMs revise essays using feedback generated by Qwen3-8B models trained on LLM-generated feedback filtered by either FeedEval or GPT-5.1. These trained models are the same ones analyzed in section 5.2.1, which are trained to jointly generate essay scores and feedback. The feedback is provided in a structured format with trait scores, as described in section 4.4. In contrast, in section 5.3.1, the small LLMs directly receive LLM-generated feedback filtered by FeedEval or GPT-5.1 without involving any intermediate model training. Figure 11 presents the results.

Overall, revisions guided by feedback generated by the model trained on FeedEval-filtered high-quality feedback exhibit trends consistent with those observed in section 5.3.1, yielding larger revision gains than feedback guided by the same model trained on low-quality feedback. By comparison, feedback provided by the model trained on GPT-5.1-selected high-quality feedback does not consistently yield larger revision gains than its low-quality counterpart. Across traits, feedback produced by the model trained on FeedEval-selected high-quality feedback results in greater revision gains than feedback generated by the model trained on GPT-5.1-selected high- or low-quality data.

In summary, consistent with the findings from essay scoring experiments, when FeedEval-selected high-quality feedback is used as a structured supervision signal together with scores, it serves as an effective training signal for essay feedback generation and leads to more meaningful essay revisions.

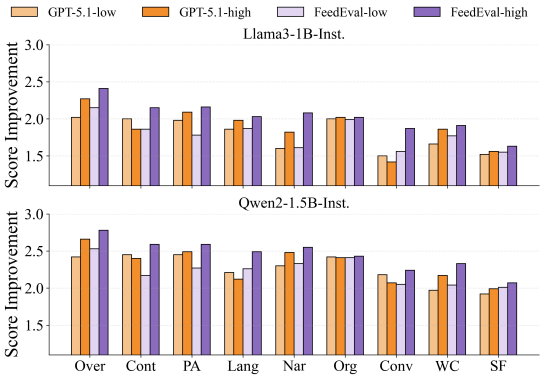

Figure 11: Average essay score improvement across traits on ASAP++ after revisions guided by models trained on feedback labels of high- and low-quality identified by FeedEval and GPT-5.1.

K Additional Materials

K.1 Dataset Statistic

We present a comprehensive statistical analysis of the datasets curated to train dimension-specific evaluator LLMs in Table 18. We further report statistics of FeedEval-identified high- and low-quality essay feedback on the ASAP++ dataset in Table 19.

K.2 Materials for Training Essay Evaluation Models

Table 20 and 21 demonstrates the inputs (user prompt) and outputs (assistant prompt) for training essay evaluation models.

K.3 Feedback Rating Dimensions used in Human Evaluation

In section 5.3.2, human experts rated the high- and low-quality essay feedback filtered by FeedEval using the following three dimensions in a 5-likert scale. The dimensions are adopted from Steiss et al. (2024).

- **Faithfulness to essay (D1)**: Does the feedback adequately and accurately reflect the content of the essay?
- **Usefulness for revision (D2)**: Does the feedback sufficiently address areas for improvement?
- **Rubric alignment (D3)**: Is the feedback grounded in the rubric?

Dataset	Task-type	# of Feedback Pair	Chosen Feedback Word Count (Min-Max)	Rejected Feedback Word Count (Min-Max)
Specificity (SpecEval)	Rewarding	41730	51.93 (16 - 129)	49.76 (13 - 129)
Helpfulness	Rewarding	14158	160.08 (7 - 741)	100.77 (1 - 968)
Validity	NLI	99952	104.58 (1 - 273)	104.58 (1 - 273)

Table 18: Statistics of the dimension-specific training datasets, including the training task type, number of feedback pairs, and average word counts of chosen and rejected feedback. For the validity dataset, the chosen and rejected feedback correspond to feedback labeled as entailment and contradiction, respectively.

Dataset	Word Count (Min-Max)	Specificity Score (Min-Max)	Helpfulness Score (Min-Max)	Validity Score (Min-Max)	Avg. Score (Min-Max)
High-quality	256.09 (115 - 455)	0.574 (0.0 - 1.000)	0.423 (0.0 - 0.999)	0.202 (0.084 - 0.405)	0.400 (0.207 - 0.786)
Low-quality	271.61 (158 - 394)	0.029 (0.0 - 0.354)	0.048 (0.0 - 0.405)	0.189 (0.084 - 0.404)	0.089 (0.030 - 0.195)

Table 19: Statistics of the FeedEval-identified high- and low-quality essay feedback datasets, including average word counts, FeedEval scores for specificity, helpfulness, and validity, as well as the average score across the three dimensions.

Input (Score Only)	<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are an essay evaluator. You will receive an essay and you will need to evaluate the essay of prompt 1, focusing on the following traits: [‘sentence fluency’, ‘word choice’, ‘conventions’, ‘organization’, ‘content’]. Score the essay in JSON format, using the trait names as keys, without any additional text.<|eot_id|><|start_header_id|>user<|end_header_id|> Essay: Dear @CAPS1 @CAPS2, I believe that using computers will benefit us in many ways like talking and becoming friends will others through websites like facebook and mysace. Using computers can help us find coordibates, locations, and able ourselfs to millions of information. Also computers will benefit us by helping with jobs as in planning a house plan and typing a @NUM1 page report for one of our jobs in less than writing it. Now lets go into the wonder world of technology. Using a computer will help us in life by talking or making friends on line. Many people have myspace, facebooks, aim, these all benefit us by having conversations with one another. Many people believe computers are bad but how can you make friends if you can never talk to them? I am very fortunate for having a computer that can help with not only school work but my social life and how I make friends. Computers help us with finding our locations, coordibates and millions of information online. If we didn’t go on the internet a lot we wouldn’t know how to go onto websites that @MONTH1 help us with locations and coordinates like @LOCATION1. Would you rather use a computer or be in @LOCATION3. When your supposed to be vacationing in @LOCATION2. Million of information is found on the internet. You can as almost every question and a computer will have it. Would you rather easily draw up a house plan on the computers or take @NUM1 hours doing one by hand with ugly erazer marks all over it, you are garrenteed that to find a job with a drawing like that. Also when appling for a job many workers must write very long papers like a @NUM3 word essay on why this job fits you the most, and many people I know don’t like writing @NUM3 words non-stopp for hours when it could take them I hav an a computer. That is why computers we needed a lot now adays. I hope this essay has impacted your descion on computers because they are great machines to work with. The other day I showed my mom how to use a computer and she said it was the greatest invention sense sliced bread! Now go out and buy a computer to help you chat online with friends, find locations and millions of information on one click of the button and help your self with getting a job with neat, prepared, printed work that your boss will love.<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Output (Score Only)	{‘sentence fluency’: 3.0, ‘word choice’: 4.0, ‘conventions’: 4.0, ‘organization’: 4.0, ‘content’: 4.0, ‘overall’: 9.0}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Table 20: An example of an input and its corresponding output used to train essay evaluation models to generate multi-trait scores only.

Input (Score + Feedback)	<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are an essay evaluator. You will receive an essay and you will need to evaluate the essay of prompt 1, focusing on the following traits: [‘sentence fluency’, ‘word choice’, ‘conventions’, ‘organization’, ‘content’]. Output only the literal evaluation in JSON format, using the trait names as keys, without any additional text.<|eot_id|><|start_header_id|>user<|end_header_id|> Essay: Dear @CAPS1 @CAPS2, I believe that using computers will benefit us in many ways like talking and becoming friends will others through websites like facebook and mysace. Using computers can help us find coordibates, locations, and able ourselfs to millions of information. Also computers will benefit us by helping with jobs as in planning a house plan and typing a @NUM1 page report for one of our jobs in less than writing it. Now lets go into the wonder world of technology. Using a computer will help us in life by talking or making friends on line. Many people have myspace, facebooks, aim, these all benefit us by having conversations with one another. Many people believe computers are bad but how can you make friends if you can never talk to them? I am very fortunate for having a computer that can help with not only school work but my social life and how I make friends. Computers help us with finding our locations, coordibates and millions of information online. If we didn’t go on the internet a lot we wouldn’t know how to go onto websites that @MONTH1 help us with locations and coordinates like @LOCATION1. Would you rather use a computer or be in @LOCATION3. When your supposed to be vacationing in @LOCATION2. Million of information is found on the internet. You can as almost every question and a computer will have it. Would you rather easily draw up a house plan on the computers or take @NUM1 hours doing one by hand with ugly erazer marks all over it, you are garrenteed that to find a job with a drawing like that. Also when appling for a job many workers must write very long papers like a @NUM3 word essay on why this job fits you the most, and many people I know don’t like writing @NUM3 words non-stopp for hours when it could take them I hav an a computer. That is why computers we needed a lot now adays. I hope this essay has impacted your descion on computers because they are great machines to work with. The other day I showed my mom how to use a computer and she said it was the greatest invention sense sliced bread! Now go out and buy a computer to help you chat online with friends, find locations and millions of information on one click of the button and help your self with getting a job with neat, prepared, printed work that your boss will love.<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Output (Score + Feedback)	{‘sentence fluency’: {‘score’: 3.0, ‘rationale’: “Quotations such as ‘I believe that using computers will benefit us in many ways like talking and becoming friends will others through websites like facebook and mysace.’ and ‘Computers help us with finding our locations, coordibates and millions of information online.’ demonstrate a lack of fluidity, as grammatical errors and awkward phrasing disrupt the reading. Sentences often lack variation in structure and length, contributing to a mechanical feeling rather than a natural flow. To improve, the writer should focus on varying sentence structures and ensuring grammatical accuracy to maintain a smoother rhythm and flow.”}, ‘word choice’: {‘score’: 4.0, ‘rationale’: “Expressions like ‘wonder world of technology’ and ‘greatest invention sense sliced bread’ attempt a conversational tone but come across as clichéd and vague. The repeated use of phrases like ‘millions of information’ lacks precision. More precise and varied language choices would enhance the essay’s impact, making it more engaging and specific.”}, ‘conventions’: {‘score’: 4.0, ‘rationale’: “The text contains multiple spelling and grammatical errors such as ‘coordibates,’ ‘garrenteed,’ and ‘appling,’ which detract from readability. Capitalization errors like ‘facebook’ and ‘mysace’ also occur. Careful proofreading for spelling, punctuation, and grammar, along with adherence to capitalization norms, would significantly improve clarity.”}, ‘organization’: {‘score’: 4.0, ‘rationale’: “The essay follows a basic structure with introductory, body, and concluding elements, as demonstrated in ‘I hope this essay has impacted your descion on computers.’ However, the transitions between ideas can be abrupt, and the overall organization could better support the clarity of the argument. Strengthening organization could involve developing clearer paragraph divisions and ensuring each paragraph supports a singular, coherent idea.”}, ‘content’: {‘score’: 4.0, ‘rationale’: “The content covers favorable aspects of computer use but lacks depth, as seen in the repetitive mention of ‘talking and becoming friends’ without substantial exploration of other potential benefits or drawbacks. Main ideas are present but require further development and detail to engage the reader fully. Enhancing content could involve a more balanced exploration of topics with specific examples and evidence to support claims, addressing different viewpoints for a more rounded perspective.”}, ‘overall’: {‘score’: 9, ‘rationale’: “nan”}}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Table 21: An example of an input and its corresponding output used to train essay evaluation models to jointly generate multi-trait scores and corresponding feedback.