Enhancing Pedagogical Feedback Quality of Large Language Models via Learning Analytics in a Kanban-Based Platform for Student Self-Regulated Learning Support

1st Qornain Aji
Department of Electrical and Information Engineering
Universitas Gadjah Mada
Yogyakarta, Indonesia
qornain.aji@mail.ugm.ac.id

2nd Indriana Hidayah
Department of Electrical and Information Engineering
Universitas Gadjah Mada
Yogyakarta, Indonesia
indriana.h@ugm.ac.id

3rd Syukron Abu Ishaq A.
Department of Electrical and Information Engineering
Universitas Gadjah Mada
Yogyakarta, Indonesia
syukron.abu@ugm.ac.id

***Abstract*—This study aims to enhance the quality of pedagogical feedback generated by Large Language Models (LLMs) to support student Self-Regulated Learning (SRL), addressing the scalability crisis of human expert feedback. The enhancement focuses on three core functions such as feedback, motivation, and appreciation. A comparative experimental method was employed without fine-tuning, evaluating various Context Engineering schemes enriched with Learning Analytics (LA) on two models: Llama 3.1 8B and Llama 3.3 70B. Evaluation was quantitative (BERTScore, BARTScore, LLM-as-a-Judge) and qualitatively validated by educational psychology experts.**

Results indicate that the effectiveness of the enhancement significantly depends on the model size. The smaller model (8B) achieved the highest performance increase with complex context (ReAct + LA), suggesting a reliance on rich, structured guidance to compensate for its internal limitations. Conversely, the larger model (70B), despite having a superior Baseline performance, showed degraded performance when given overly complex context. The best-performing model (Llama 8B with ReAct + LA) was validated by human experts as ”Usable with minor revision”. The study concludes that LA-enriched context engineering is an effective, practical, and computationally efficient strategy for improving pedagogical feedback, particularly for smaller LLMs.

***Index Terms*—Large Language Model, Self-Regulated Learning, Learning Analytics, Context Engineering, Pedagogical Feedback**

I. Introduction

The current landscape of higher education demands that s tudents possess the ability to learn independently. However, many are unprepared to face these demands. This creates a serious gap in *Self-Regulated Learning* (SRL) skills, which can hinder learning success and disrupt equitable access to education [1], [2]. This challenge is exacerbated by the digital technologies that characterize modern learning. While offering flexibility, these technologies also present significant distractions and require greater self-discipline [3]. Institutional failure to provide adequate support for SRL development is not merely an academic issue; it is a social issue that impacts students' long-term career prospects and the overall effectiveness of the education system [1].

To address the SRL gap, Zimmerman proposed a framework encompassing the phases of planning (*forethought*), execution (*performance*), and self-reflection (*self-reflection*), wherein SRL is a combination of cognitive, motivational, and behavioral strategies [1], [4]. Individuals with effective self-regulation capabilities are formed through goal setting, strategy planning, active self-monitoring, and thoughtful self-evaluation. A substantial body of research confirms a strong relationship between these skills and academic achievement [1], [2]. However, their development is often hindered by personal, contextual, and social factors, ranging from poor time management and digital distractions to low metacognitive awareness. Metacognition itself refers to the knowledge, awareness, and regulation of one's own thinking, playing a crucial role in building students' awareness of their learning processes [5]. These barriers necessitate structured interventions to guide students through this complex process [3], [6].

The most effective traditional method for developing SRL skills is personalized feedback from human experts. High-quality feedback that focuses not only on the quantity of completed tasks but also on the student's learning process and self-regulation serves as a powerful catalyst for improvement [15, 16] [7], [8]. However, this ”gold standard” faces a scalability crisis; increasing student numbers and diverse support needs make the personalized guidance model difficult to implement on a large scale. This is reinforced by studies showing that budgetary pressures drive larger class sizes, thereby severely limiting personal interaction between lecturers and students [9], [10]. This creates an urgent need for innovative technology-based solutions that can complement human expertise.

One such technology is the use of *Large Language Models* (LLMs). LLMs can be employed as virtual tutors capable of providing fast and consistent feedback. LLMs have been utilized in various educational research contexts with diverse breakthroughs. Recent findings show that when LLMs are used for assessment and feedback, *artificial intelligence* (AI) tends to be generous in grading, resulting in scores higher than those given by human experts. Meanwhile, peer and lecturer assessments tend to be lower and more aligned with student performance. AI feedback is often structured but still requires human expert oversight to remain pedagogically relevant, leading to the recommendation of a hybrid approach (combining AI and human experts) [11]. A hybrid approach is recommended as it can combine the speed and scalability of AI with the contextual mastery, empathetic understanding, and pedagogical nuances of human assessment [11]. This recommendation underscores that LLM responses still require expert supervision to be relevant. On the other hand, efforts to standardize AI-tutor evaluation indicate that although current models like GPT-4 are strong in answering questions, they tend to provide answers to students too quickly and lack process guidance. This fact suggests that LLMs are not yet ideal as tutors and cannot yet replace human experts in assessing pedagogical quality [12].

Therefore, this thesis focuses on improving the performance quality of LLM responses as feedback providers to approach the quality of human experts, both quantitatively and qualitatively. This improvement is executed through an evaluation framework that combines human expert references as *ground truth* with quality assessments evaluating relevant pedagogical dimensions [11], [12]. To support this measurement process, this thesis utilizes structured learning process traces. Student learning data, presented in the form of a Kanban board, is used solely as an operational tool to organize and extract process data (card movement, checklists, completion time) that can be mapped to SRL phases (planning-monitoring-reflecting) [13]. The use of the Kanban board serves only as a data enrichment instrument to ensure the evaluation and improvement of LLM feedback quality proceeds more purposefully. In this research, as explained in the subsequent chapter, the focus of novelty lies in the series of evaluations and refinements of the LLM *prompts*. This research aims to improve three main pedagogical functions—feedback, motivational support, and appreciative support—with the practical goal of matching human expert quality on quantitative metrics while simultaneously meeting qualitative quality standards [7], [8], [11], [12]. By comparing LLM outputs with expert specialists, this study assesses the improvement in LLM performance for consideration as a massive yet personalized complementary pedagogical agent for students.

II. Related Works

Woodrow, Koyejo, and Piech [14] address the *Feedback Alignment Challenge*, namely the tendency of LLM-generated feedback to be generic and misaligned with course-specific rigor, terminology, and instructor preferences. They propose *Direct Preference Optimization* (DPO) with teachers in the loop, where TAs select or edit preferred outputs from model pairs during grading, producing preference data that enables fine-tuning without a reward model as in RLHF. Their three-stage pipeline (preference collection, DPO training, inference) forms a self-improving system that increasingly aligns with course expectations across assignments. Using Llama 3.1 8B fine-tuned with HuggingFace DPOTrainer, they conducted controlled blind evaluations with experts and real deployments in large university courses, supported by automated assessment frameworks (e.g., Scarlatos et al. [15]). Results show that DPO outperformed GPT-4o in preference (56.8% vs. 40.2%), insightfulness, correctness, tone positivity, and consistency, though GPT-4o remained stronger in conciseness and suggestions. Deployment demonstrated feasibility at scale and potential for fairness monitoring, positioning DPO as a practical strategy for course-aligned AI feedback generation.

Jacobsen and Weber [16] examine how prompt quality shapes LLM feedback effectiveness in higher education and compare LLM-generated feedback with novice and expert human feedback. They develop a theory-based prompt manual covering role/context specification, mission clarity, instructional logic, and domain specificity, then craft low-, medium-, and high-quality prompts to systematically test prompt effects using ChatGPT-4. Across a geometry learning task involving typical student misconceptions, 60 LLM outputs were coded with pedagogical criteria (e.g., concreteness, activation, empathy, specificity) and compared with feedback written by preservice teachers and experts using the same high-quality prompt. Results show that only high-quality prompts reliably produce strong LLM feedback; LLMs outperform novices in most categories and even exceed experts in explanation, guiding questions, and specificity, while affective dimensions remain mixed. Efficiency gains were substantial, as ChatGPT-4 produced dozens of feedback instances in the time an expert crafted one. The authors highlight the necessity of structured prompt design training for educators and note limitations such as hallucination risks and the need to test student perceptions and transferability across contexts.

Ouyang et al. [17] introduce an RLHF pipeline to align language models with user intent, addressing limitations of large pretrained models that may generate incorrect, unhelpful, or unsafe outputs despite high linguistic proficiency. Their three-stage approach includes supervised fine-tuning on human demonstrations, reward model (RM) training from human comparisons, and policy optimization with PPO using RM predictions as rewards. Drawing on prompts from labelers and real API usage, they train reward models (initialized from GPT-3 6B) and RLHF policies with careful stability controls (KL penalties, learning rate warmup, single-epoch RM training to avoid overfitting). The resulting InstructGPT models show substantial alignment gains: the 1.3B RLHF model is preferred over the much larger GPT-3 175B baseline, while exhibiting improved truthfulness, reduced toxicity, and minimal regression on NLP benchmarks. The study demonstrates that human feedback alignment can outperform mere parameter scaling, while acknowledging remaining errors and offering a blueprint for instruction-following models widely adopted today.

Li and Ma [18] propose CodeRunner Agent, an AI-powered system integrated into Moodle to support self-regulated learning (SRL) in programming education, addressing limitations of external LLM tools that lack course context and produce feedback disconnected from curricula or student behavioral data. The system combines a lecture viewer, CodeRunner execution/grading, and xAPI-based learning analytics to capture rich process data. Its pedagogical foundation is the PPESS model (Planning, Program creation, Error correction, Self-monitoring, Self-reflection), allowing students to request AI help targeted to their current SRL phase. Two context engines drive feedback relevance: LACE (Learning Analytics Context Engine) summarizes engagement and performance patterns from logs, while KCE (Knowledge Context Engine) manages curated course materials, concepts, solution structures, and typical errors. These engines inject SRL and knowledge-context cues into prompts to ensure feedback aligns with curriculum goals while avoiding spoon-feeding. The paper presents a design and planned evaluations—including behavioral analytics and qualitative assessments—to validate impact on SRL, code performance, and pedagogical safety, highlighting the feasibility of an end-to-end, context-aware, LMS-native AI support system.

Strickroth, Kreidenweis, and Götzfried propose a real-time, networked extension of AgileBoard4Teaching to address orchestration challenges in collaborative, task-based classrooms where teachers struggle to monitor diverse group progress simultaneously. Their client-server system includes an authoring mode for task setup, a Kanban interface for students, and a teacher dashboard that displays per-group progress summaries, pending reviews, hand-raise signals, timers, messaging, and live synchronization via WebSockets, with storage handled through Java Servlets and MariaDB. Two evaluations were conducted: a field study with 8th-grade students using a within-subjects comparison of offline versus networked versions, and a simulation study with experienced teachers. Findings show strong preference for the networked system, improved workflow fluency, high usability scores (student SUS 84, teacher SUS 92), and high engagement observed in interaction logs; the simulation study also yielded positive reception despite some bugs. Teachers valued real-time monitoring and ease of scaling boards, suggesting refinements for large-class readability and analytics features. Overall, the study demonstrates the feasibility and pedagogical value of real-time dashboards for classroom orchestration.

In light of the comparative review, this study adopts a pragmatic hybrid approach that prioritizes resource-efficient context alignment, namely high-quality prompt engineering and the injection of learning-analytics signals from Kanban process data, while maintaining strong pedagogical oversight. Instead of relying on costly full-model retraining, the method enriches teacher-authored prompts with contextual features such as task states, time-on-task, and review requests to make LLM feedback more course and SRL aware. Preference-based fine-tuning approaches (e.g., DPO or RLHF) are considered only as long-term options when sufficient preference labels and computational resources are available. Evaluation will combine quantitative comparisons of LLM outputs against expert-based rubrics and automated metrics with qualitative expert judgment to ensure pedagogical soundness. The accompanying real-time dashboard further enables continuous monitoring and iterative refinement of prompts and contextual cues. Altogether, this design aims to deliver a scalable, low-cost pipeline that leverages prompt/context engineering and learning-analytics signals to generate specific, actionable, and pedagogically aligned feedback while keeping instructors firmly in the loop.

III. Methodology

The research was conducted through several stages. The research starting with a comprehensive literature review to understand the current methods for improving LLM performance in producing higher feedback quality. This review covered techniques such as Reinforcement Learning from Human Feedback (RLHF), Direct Preference Optimization (DPO), and Context Engineering, as well as the use of Learning Analytics (LA) to enrich context for LLMs. The literature review also examined the pedagogical aspects of feedback in supporting Self-Regulated Learning (SRL) and the use of Kanban-based platforms for organizing student learning processes. After identifying the research gap, the next step involved selecting various LLM models and prompting techniques for evaluation.

A. Tools

The tools used in this final project consist of both hardware and software as supporting resources. The hardware employed is personally owned. The function of these hardware and software tools is to facilitate code development, chatbot implementation, quantitative and qualitative evaluations, as well as dataset collection. The following is the list of tools used in this research project.

1) **Hardware:**
   - MacBook Pro 14 inch: M1 Pro 10-core processor @ 3.2 GHz
   - Internal GPU: M1 Pro 16-core GPU
   - Neural Engine: 32-core
   - RAM: 16 GB Unified Memory (200 GB/s memory bandwidth)
   - Storage: 1 TB internal Solid State Drive (macOS 26 Tahoe)
2) **Software:**
   - Visual Studio Code version 1.105.1: Used for writing and managing code.
   - Google Docs: Web-platform for drafting questionnaires, performing revisions, and documenting discussion results.
   - Google Sheets: Web-platform that is used to visualize and manage student Kanban learning data to share with experts for completion and assessment, facilitating collaboration and metric collection.
   - Anaconda: Python package and virtual environment manager used during development.

3) **Key Libaries and Frameworks:**

- `Python 3.12.7`: Primary programming language for developing the LLM-based chatbot.
- `LangChain`: Framework for building LLM applications, providing modular components for prompt orchestration, data flows, and external data integration.
- `LangGraph`: An extension of LangChain for constructing more complex application flows as graphs.
- `bert_score`: Quantitative evaluation framework to measure semantic similarity (Precision, Recall, F1) between model responses and ground truth.
- `BARTScorer`: Alternative quantitative evaluation framework (using mBART) to compute F1, Precision, and Recall based on token log-probabilities.
- `pandas`: Library for data manipulation, aggregation, and presentation of evaluation summary tables (e.g., BARTScore and GPT-4o Judge).

B. Datasets Creation

This study necessitated a specialized dataset capable of reflecting student learning processes within a **Kanban-based learning analytics platform**, while simultaneously possessing high-quality pedagogical ground truth. Due to the unavailability of public datasets meeting these specific criteria, a comprehensive methodology for **synthetic data generation** was designed. The objective was to produce a realistic and diverse dataset comprising two main components such as **student digital trace data in JSON format** representing Kanban board activities, and **aligned human expert feedback**.

The data generation process began with a deconstruction of the **Gamatutor** platform's database schema to identify key attributes reflecting SRL processes, such as column movements, timestamps, and checklist progress. To capture a broad spectrum of student engagement, five distinct behavioral profiles were defined: **"Lazy"**, **"Slightly Lazy"**, **"Average"**, **"Diligent"**, and **"Very Diligent"**. These profiles were operationalized into strict case study rules governing variables like total card count and checklist completion percentages. The content domain was restricted to **Object-Oriented Programming (OOP)** with five specific subtopics.

These structural and behavioral constraints were transformed into highly structured prompts using advanced engineering techniques, including **Role-play**, **Few-shot (providing JSON examples)**, and **Chain-of-Thought (CoT)** to guide the LLM's reasoning. Using the **ChatGPT-5 Thinking model**, **40 unique Kanban datasets** were generated. The following table summarizes the distribution of datasets across the defined behavioral profiles is shown in Table I.

Subsequently, to establish the **ground truth**, the raw **JSON data** was transformed into **human-readable tabular formats**. These tables were presented to **three human experts** (university lecturers), who provided **structured pedagogical feedback** covering three primary functions: **feedback, motivation**, and **appreciation**. This collection of expert feedback served as the definitive **ground truth** for evaluating the LLM agents in this study.

TABLE I: Dataset summary by behavioral profile

Behavioral Profile	Total Datasets
Lazy	5
Slightly Lazy	10
Average	10
Diligent	10
Very Diligent	5
Total	**40**

C. Model Selections

This study utilizes two distinct models from the Llama family, accessible via the Groq API, Llama 3.1 8B Instant and Llama 3.3 70B Versatile, to generate pedagogical feedback supporting student SRL. The selection of these specific LLMs was driven by several strategic considerations rather than being the primary object of the research itself.

Firstly, the decision was significantly influenced by financial and computational resource constraints, a valid justification in academic educational research. The Groq API provides these models at cost-effective rates, enabling broader accessibility for research purposes [19]. The specific pricing and specifications for the models used, alongside others in the ecosystem, are detailed in Table II.

TABLE II: Groq API Model Pricing and Specifications

Model (Configuration)	TPS	Input ($/1M tokens)	Output ($/1M tokens)
Llama 4 Scout (17Bx16E)	594	$0.11	$0.34
Llama 4 Maverick (17Bx128E)	562	$0.20	$0.60
Llama Guard 4 12B	325	$0.20	$0.20
Qwen3 32B 131k	662	$0.29	$0.59
Llama 3.3 70B Versatile	**394**	**$0.59**	**$0.79**
Llama 3.1 8B Instant	**840**	**$0.05**	**$0.08**

Source: Adapted from Groq API Pricing [19]

As shown in Table II, the 8B model operates at approximately USD 0.05 for input and USD 0.08 for output per million tokens, offering high throughput (840 TPS) [19]. Meanwhile, the 70B model costs USD 0.59 for input and USD 0.79 for output per million tokens. While Llama 3.3 70B is not the least expensive option available, it offers a reasonable price-to-performance ratio for a large-scale model suitable for this study's requirements.

Secondly, the selection aligns with the perspective of Steinert et al., who emphasize the importance of open-source platforms in making LLM technology affordable and inclusive [20]. Choosing Llama models over high-cost proprietary alternatives promotes cost efficiency, which is crucial for facilitating empirical research in educational practices that often operate under resource constraints.

Finally, there is a lack of comparative studies examining the performance differences between small (8B) and large (70B)

LLMs specifically in the context of providing pedagogical feedback for Self-Regulated Learning (SRL). Investigating these differences is essential, as model size directly correlates with computational resource demands, operational costs, and the adaptability of the generated feedback. Therefore, this selection allows the study to address the research question regarding the influence of model capacity on the quality of pedagogical feedback.

D. Feedback Improvement Methods

This section details the methodologies employed to enhance the quality of pedagogical feedback generated by **LLMs** aiming to approach the standard of **human experts**. The improvement process is not reliant on a single technique but rather a **systematic multifaceted approach**. This methodology is founded on **three main pillars** which are **context engineering**, the integration of **descriptive learning analytics**, and the design of **prompt variation experiments** to test the impact of each treatment.

The first pillar is context engineering which focuses on designing a theoretically robust **Baseline prompt**. This prompt is crafted to fulfill specific pedagogical objectives such as providing adaptive guidance and enhancing self efficacy grounded in educational psychology literature. To ensure structure and consistency the prompt is divided into five logical components **Role Definition, Internal Logic, Contextual Information, Assessment Prompt, and Input Output Format**. The quality of this prompt architecture was validated using a theory driven prompt manual where the prompt used in this study achieved a score of **13 out of 16** classifying it as High quality.

The second pillar involves the integration of **learning analytics** to inject situational awareness into the model. This approach transforms raw JSON data from student Kanban boards into actionable insights. By applying **Descriptive Analytics** a series of 12 key indicators such as `aging_wip_h`, `stuck` and `alerts` are extracted from the board data. the Fig 1 illustrates the process of

These indicators are based on established Kanban management practices and calibrated with a **2 hour threshold** relevant to the research case study. This analytic process generates a concise analytic summary which is then inserted into the prompt to enable the LLM to provide personalized data driven feedback.

The third pillar is the creation of **prompt variations** to empirically measure the effectiveness of the first two pillars. A systematic experiment was designed to compare four different prompt conditions. These variations include the **Baseline Prompt**, the Prompt with added **Material List** context, the Prompt enriched with **learning analytics** data, and the Prompt combining **learning analytics** with a **ReAct (Reasoning plus Acting)** mechanism. The **ReAct mechanism**, implemented using the **LangGraph** framework, is designed to force the model to perform internal reasoning on analytic data before generating output. All these experiments are conducted on two LLM models of different sizes (Llama 3.1 8B Instant and Llama 3.3 70B Versatile), to analyze the quantitative and qualitative impact of each strategy on the quality of pedagogical feedback produced.

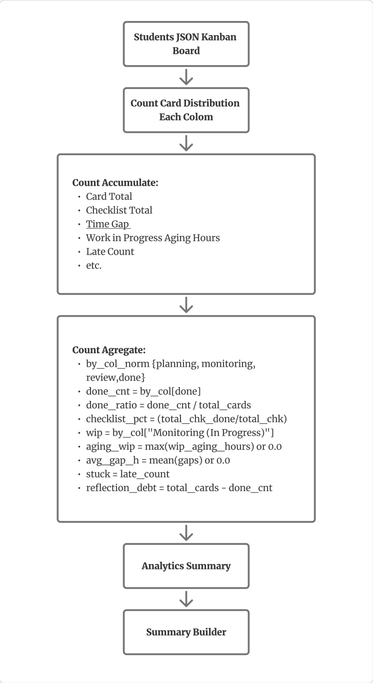
Students JSON Kanban Board
Count Card Distribution Each Colom
Count Accumulate:
Card Total
Checklist Total
Time Gap
Work in Progress Aging Hours
Late Count
etc.
Count Agregate:
by_col_norm {planning, monitoring, review,done}
done_cnt = by_col[done]
done_ratio = done_cnt / total_cards
checklist_pct = (total_chk_done/total_chk)
wip = by_col["Monitoring (In Progress)"]
aging_wip = max(wip_aging_hours) or 0.0
avg_gap_h = mean(gaps) or 0.0
stuck = late_count
reflection_debt = total_cards - done_cnt
Analytics Summary
Summary Builder

Fig. 1: Descriptive Analytics Process for Kanban Data

E. Evaluation Methods

The evaluation strategy is designed to systematically assess the effectiveness of the proposed improvement methods which consist of context engineering and learning analytics followed by the creation of prompt variations. This study employs a comparative experimental design to evaluate the quality of pedagogical feedback generated by LLMs across these different schemes. The assessment is conducted using a dual approach comprising both quantitative and qualitative evaluations to ensure a comprehensive analysis of the model outputs.

Quantitative evaluation focuses on measuring semantic similarity and relevance using advanced metrics based on contextual embeddings. The primary metrics employed include BERTScore [21] and BARTScore [22] which calculate the semantic alignment between the LLM generated responses and the human expert ground truth. Additionally the LLM as a Judge [23] method utilizes the GPT 4o model to evaluate the feedback based on specific pedagogical dimensions such as feedback quality motivation support and appreciation support. This structured automated assessment provides objective data regarding the precision and recall of the generated feedback.

Qualitative evaluation serves as the gold standard for validating the pedagogical value of the feedback. This process involves a direct review by an educational psychology expert who assesses the output using a comprehensive rubric grounded in established theoretical frameworks including Self

Determination Theory [24], Positve Psychology [25] and the feedback models of Hattie and Timperley [26]. The expert evaluates the feedback using a Likert scale questionnaire to measure its effectiveness in supporting student Self Regulated Learning ensuring the responses are not only technically accurate but also psychologically supportive and educationally relevant.

IV. Result and Analysis

This chapter presents the results and analysis of the experiments which were conducted by implementing the improvement methods outlined in the previous chapter. The focus is on evaluating the effectiveness of context engineering and learning analytics integration in enhancing the quality of pedagogical feedback generated by LLMs. The results are analyzed based on quantitative metrics such as BERTScore, BARTScore, and LLM-as-a-Judge evaluations, as well as qualitative assessments by educational psychology experts.

The comparative analysis of feedback improvement methods reveals distinct performance patterns dependent on model size and context complexity. Quantitative evaluation using BERTScore and BARTScore indicates that the smaller 8B model achieved the most significant quality improvement when integrating complex context engineering and Learning Analytics, specifically the ReAct plus LA variation. This suggests that smaller models rely heavily on structured guidance to enhance semantic quality [27], [28]. The summary of BERTScore and BARTScore results can be seen in Table III and Table IV.

TABLE III: Summary of BERTScore Results per Context-Engineering Variation

Variant	**8B (Llama 3.1 8B)**			**70B (Llama 3.3 70B)**		
	Prec.	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
Baseline	0.6505	0.6719	0.6608	0.6567	0.6815	0.6687
Material List	0.6500	0.6700	0.6597	0.6620	0.6805	**0.6710**
LA	0.6586	0.6733	0.6657	0.6529	0.6774	0.6648
ReAct + LA	0.6613	0.6716	**0.6663**	0.6561	0.6760	0.6658

TABLE IV: Summary of BARTScore Results per Context-Engineering Variation

Variant	**8B (Llama 3.1 8B)**			**70B (Llama 3.3 70B)**		
	Prec.	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
Baseline	-11.09	-12.36	-11.72	-11.16	-12.21	-11.69
Material List	-11.17	-12.53	-11.85	-11.03	-12.18	**-11.61**
LA	-10.90	-12.18	-11.54	-11.25	-12.13	-11.69
ReAct + LA	-10.58	-12.23	**-11.40**	-10.89	-12.45	-11.67

Conversely, the larger 70B model demonstrated optimal performance with simpler context engineering such as the Material List variation, while complex contexts like ReAct plus LA tended to stagnate or degrade its performance [27], [28].

Regarding the influence of model size, the analysis shows that the 70B model possesses a superior baseline capacity compared to the 8B model, particularly in relevance metrics measured by LLM-as-a-Judge (see Table V and Table VI).

TABLE V: Summary of Overall LLM-as-a-Judge Results per Method

Method	**8B (Llama 3.1 8B)**			**70B (Llama 3.3 70B)**		
	Prec.	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
Baseline	1.0000	0.4567	0.5371	1.0000	0.6417	0.7271
Material List	0.9333	0.4767	0.5627	1.0000	0.6600	**0.7571**
LA	1.0000	0.5017	0.5896	0.9333	0.5717	0.6399
ReAct + LA	1.0000	0.5217	**0.6121**	1.0000	0.5883	0.6715

TABLE VI: Baseline Performance Comparison Between 8B and 70B Models

Metric (F1-Score)	**8B Model**	**70B Model**	**Winner**
BERTScore	0.6608	0.6687	70B
BARTScore	-11.7282	-11.6905	70B
LLM-as-a-Judge	0.5371	0.7271	70B

However, the 8B model exhibited higher responsiveness to context engineering interventions. The performance comparison across different metrics and models is visualized in Figure 2. Qualitative analysis highlights that the 70B model generates more prescriptive and linguistically varied feedback, whereas the 8B model tends to be descriptive and repetitive, often failing to separate appreciation from critique.

The validation by human experts confirms the potential of the system, with the best-performing model (Llama 3.1 8B with ReAct plus LA) being rated as ”Usable with minor revision.” Experts assessed the feedback, motivation, and appreciation aspects positively, as detailed in Tables VII, VIII, and IX.

TABLE VII: Expert Assessment on Feedback Quality Aspects

Indicator	**Question**	**Score**
F1 (Clarity)	Feedback is easy to understand.	4
F2 (Specificity)	Feedback gives specific suggestions.	4
F3 (Relevance)	Feedback fits the learning activity.	3
F4 (Constructiveness)	Feedback helps fix errors.	4
F5 (Feed-forward)	Feedback provides next steps.	4

TABLE VIII: Expert Assessment on Motivational Support Aspects

Indicator	**Question**	**Score**
M1 (Intrinsic Drive)	Makes me more excited to learn.	4
M2 (Goal Drive)	Helps stay focused on learning goals.	5
M3 (Autonomy)	Gives freedom in learning methods.	4
M4 (Competence)	Makes me feel more confident.	4
M5 (Relatedness)	Support feels empathetic.	3

Experts noted areas for improvement such as the need for more specific context references and a less robotic tone in motivational support. These findings also reinforce the evaluation of prompt quality based on the theory-driven prompt manual, where a high prompt score does not necessarily guarantee optimal model output without considering model size and capabilities [16]. The feedback was deemed effective in maintaining focus on learning goals although it sometimes lacked the empathetic connection found in human interaction.

V. Conclusions

This study conducted experiments to design improve and evaluate the quality of pedagogical feedback generated by

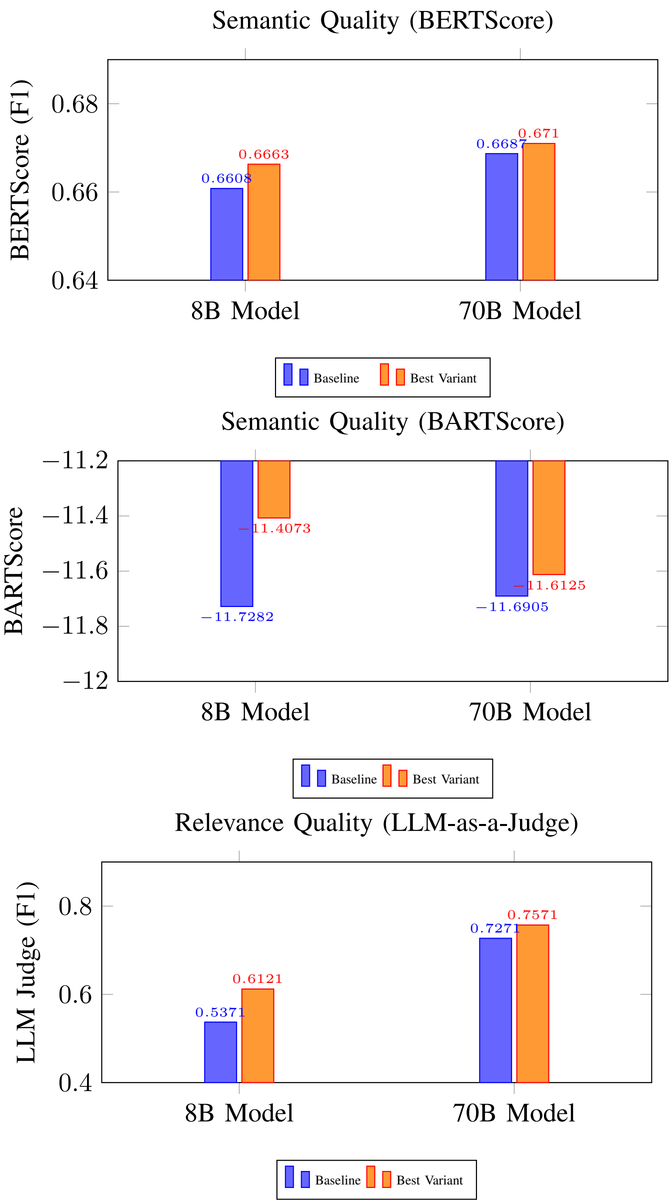

Fig. 2: Performance Comparison of 8B vs 70B Models across Different Evaluation Metrics.

TABLE IX: Expert Assessment on Appreciation and Emotional Support

Indicator	Question	Score
A1 (Effort Recog.)	Appreciates effort, not just results.	5
A2 (Proportional)	Praise feels sincere and appropriate.	3
A3 (Confidence)	Appreciation boosts confidence.	4
A4 (Emotional Supp.)	Provides calming responses.	4
A5 (Positive Reinf.)	Language makes me feel valued.	3

LLMs focusing on enhancing feedback motivation and appreciation to support student Self Regulated Learning. The improvement was achieved through context engineering methods enriched with learning analytics extracted from synthetic student Kanban board data without fine tuning the models. Based on the analysis the following conclusions are drawn to address the research questions.

First the method of context engineering combined with learning analytics proved successful in improving LLM feedback quality compared to the Baseline although its effectiveness depended heavily on the model size. For the small model Llama 3.1 8B the addition of complex context involving ReAct and learning analytics consistently provided the highest performance improvements across all quantitative metrics. This indicates that small models benefit significantly from rich and structured context guidance to produce feedback that is more relevant and of higher semantic quality.

Second model size has a significant and bidirectional influence on feedback quality. The large model Llama 3.3 70B showed far superior Baseline performance compared to the 8B model particularly in terms of relevance. However the 70B model did not benefit from complex contexts achieving its best performance with simpler contexts such as the Material List. The addition of complex ReAct and LA contexts even degraded its performance likely because the prompt overly constrained the large model forcing it to mimic the descriptive style of the small model rather than approaching the ground truth.

Third based on human expert perception the feedback generated by the best performing model Llama 3.1 8B with ReAct and LA was declared Usable with minor revision. Experts confirmed the potential of the chatbot as an accessible scaffolding system. Suggested revisions focused on adjusting language tonality eliminating negative diction addressing response category overlap improving feedback specificity and increasing the variation of motivational responses. These findings also emphasize that a good score from a theory driven prompt manual does not necessarily guarantee optimal model output without considering other factors like model size and language capabilities.

In conclusion this research determines that context engineering enriched with learning analytics is an effective practical and computationally efficient strategy for improving pedagogical feedback especially when applied to small LLM models.

VI. Future Works

Based on the limitations and potential developments identified in this study future research can consider several advancements to further enhance the quality and applicability of pedagogical feedback agents.

First further investigation is needed regarding the influence of complex context engineering on large LLMs to understand their boundaries and characteristics in responding to rich contexts. Understanding why larger models may degrade with complex prompting is crucial for optimizing their use in educational settings.

Second developing human expert evaluation using A B testing is recommended where experts assess two feedback variations to directly select the best one allowing for a more comparative and definitive assessment of quality.

Third experiments should include more diverse prompt variations including those adjusting language tonality according to user demographic characteristics to improve the empathetic connection and relevance perceived by students.

Fourth further research is required to identify optimal prompts for generating high quality feedback going beyond assessments based solely on theory driven prompt manuals. This involves exploring prompt structures that yield the best empirical results rather than just theoretical compliance.

Fifth conducting statistical hypothesis testing is essential to strengthen the validity of quantitative findings and evolve the current exploratory direction into robust hypotheses.

Sixth evaluating LLMs with superior Indonesian language performance is suggested to observe the impact on feedback quality as language proficiency significantly affects the nuance and accuracy of pedagogical responses.

Finally using real student learning data to extract learning analytics and involving a larger number of experts for validation and ground truth collection will enhance the accuracy and relevance of research results ensuring the system is robust for real world application.

References

[1] C. Liu, Z. A. Bakar, and Q. Xu, “Self-regulated learning and academic achievement in higher education: A decade systematic review,” *International Journal of Research and Innovation in Social Science (IJRISS)*, vol. 9, no. 03, pp. 4488–4504, 2025. [Online]. Available: https://dx.doi.org/10.47772/IJRISS.2025.90300358

[2] S. Xiao, K. Yao, and T. Wang, “The relationships of self-regulated learning and academic achievement in university students,” *SHS Web of Conferences*, vol. 60, p. 01003, 2019, pHECSS2018. [Online]. Available: https://doi.org/10.1051/shsconf/20196001003

[3] E. Chitra, N. Hidayah, M. Chandratilake, and V. D. Nadarajah, “Self-regulated learning practice of undergraduate students in health professions programs,” *Frontiers in Medicine*, vol. 9, p. 803069, 2022. [Online]. Available: https://doi.org/10.3389/fmed.2022.803069

[4] A. Akdeniz, “Exploring the impact of self-regulated learning intervention on students’ strategy use and performance in a design studio course,” *International Journal of Technology and Design Education*, pp. 1–35, 2022, advance online publication. [Online]. Available: https://doi.org/10.1007/s10798-022-09798-3

[5] B. J. Zimmerman and A. R. Moylan, “Self-regulation: Where metacognition and motivation intersect,” in *Handbook of Metacognition in Education*, D. J. Hacker, J. Dunlosky, and A. C. Graesser, Eds. New York, NY: Routledge, 2009, pp. 299–315.

[6] F. Nyman, ““you’re not learning skills—you’re just realizing what you can do”: a preliminary study of self-regulation in higher education,” *Frontiers in Education*, vol. Volume 9 - 2024, 2024.

[7] L. F. Yang, Y. Liu, and Z. Xu, “Examining the effects of self-regulated learning-based teacher feedback on english-as-a-foreign-language learners’ self-regulated writing strategies and writing performance,” *Frontiers in Psychology*, vol. Volume 13 - 2022, 2022.

[8] T. Bock, E. Thomm, J. Bauer, and B. Gold, “Fostering student teachers’ research-based knowledge of effective feedback,” *European Journal of Teacher Education*, vol. 47, no. 2, pp. 389–407, 2024. [Online]. Available: https://doi.org/10.1080/02619768.2024.2338841

[9] H. E. Schaffer, K. R. Young, E. W. Ligon, and D. D. Chapman, “Automating individualized formative feedback in large classes based on a directed concept graph,” *Frontiers in Psychology*, vol. 8, p. 260, 2017, article 260. [Online]. Available: https://doi.org/10.3389/fpsyg.2017.00260

[10] A. Pardo, J. Jovanović, S. Dawson, D. Gašević, and N. Mirriahi, “Using learning analytics to scale the provision of personalised feedback,” *British Journal of Educational Technology*, vol. 50, no. 1, pp. 128–138, Jan. 2019. [Online]. Available: https://doi.org/10.1111/bjet.12592

[11] M. Usher, “Generative ai vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education,” *Assessment & Evaluation in Higher Education*, 2025, published online: Apr. 9, 2025. [Online]. Available: https://doi.org/10.1080/02602938.2025.2487495

[12] K. K. Maurya, K. V. A. Srivatsa, K. Petukhova, and E. Kochmar, “Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors,” *arXiv preprint arXiv:2412.09416*, 2025, version 2. [Online]. Available: https://arxiv.org/abs/2412.09416

[13] S. Strickroth, M. Kreidenweis, and Z. Wurm, “Learning from agile methods: Using a kanban board for classroom orchestration,” in *Proceedings of the 25th International Conference on Interactive Collaborative Learning (ICL 2022)*, ser. Lecture Notes in Networks and Systems, M. E. Auer, W. Pachatz, and T. Rüütmann, Eds., vol. 633. Springer, 2022, pp. 68–79.

[14] J. Woodrow, S. Koyejo, and C. Piech, “Improving generative ai student feedback: Direct preference optimization with teachers in the loop,” in *Proceedings of the 2025 International Conference on Educational Data Mining (EDM 2025)*. International Educational Data Mining Society, 2025, p. —. [Online]. Available: https://educationaldatamining.org/EDM2025/proceedings/2025.EDM.short-papers.166/index.html

[15] A. Scarlatos, D. Smith, S. Woodhead, and A. Lan, “Improving the validity of automatically generated feedback via reinforcement learning,” in *Artificial Intelligence in Education*, A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt, Eds. Cham: Springer Nature Switzerland, 2024, pp. 280–294.

[16] L. J. Jacobsen and K. E. Weber, “The promises and pitfalls of large language models as feedback providers: A study of prompt engineering and the quality of ai-driven feedback,” *AI*, vol. 6, no. 2, 2025. [Online]. Available: https://doi.org/10.3390/ai6020035

[17] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, “Training language models to follow instructions with human feedback,” 2022. [Online]. Available: https://arxiv.org/abs/2203.02155

[18] H. Li and B. Ma, “Design of ai-powered tool for self-regulation support in programming education,” 2025. [Online]. Available: https://arxiv.org/abs/2504.03068

[19] “Groq API Pricing,” Groq. [Online]. Available: https://groq.com/pricing

[20] S. Steinert, K. E. Avila, S. Ruzika, J. Kuhn, and S. Küchemann, “Harnessing large language models to develop research-based learning assistants for formative feedback,” *Smart Learning Environments*, vol. 11, no. 1, p. 62, 2024. [Online]. Available: https://slejournal.springeropen.com/articles/10.1186/s40561-024-00354-1

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” in *International Conference on Learning Representations (ICLR)*, 2020, also available as arXiv:1904.09675.

[22] W. Yuan, G. Neubig, and P. Liu, “Bartscore: Evaluating generated text as text generation,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 27 263–27 277.

[23] C. Koutchéme, N. Dainese, S. Sarsa, A. Hellas, J. Leinonen, and P. Denny, “Open source language models can provide feedback: Evaluating llms’ ability to help students using gpt-4-as-a-judge,” 2024.

[24] E. L. Deci and R. M. Ryan, *Intrinsic Motivation and Self-Determination in Human Behavior*. New York: Plenum, 1985.

[25] B. L. Fredrickson, “The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions,” *American Psychologist*, vol. 56, no. 3, pp. 218–226, 2001.

[26] J. Hattie and H. Timperley, “The power of feedback,” *Review of Educational Research*, vol. 77, no. 1, pp. 81–112, 2007.

[27] Y. Tang, D. Tuncel, C. Koerner, and T. Runkler, “The Few-shot Dilemma: Over-prompting Large Language Models,” *arXiv preprint arXiv:2509.13196*, Sep 2025.

[28] L. Pawlik, “How the choice of llm and prompt engineering affects chatbot effectiveness,” *Electronics*, vol. 14, no. 5, 2 2025.