Pedagogical Mediation Through Prompt Engineering: An Expert Evaluation of AI-Generated Feedback on Islamic-Integrated EFL Argumentative Writing

Sari Dewi Noviyanti, Rudi Hartono*, Hendi Pratama, Seful Bahri
English Language Education Department, Semarang State University, Semarang, Central Java 50229, Indonesia

***Abstract*—This research tested whether prompt engineering could act as a type of pedagogical mediation to enhance the quality of AI-generated student feedback on EFL students' argumentative essays in an Islamic education system. An initial pool of eight expert raters (four primary raters and four inter-raters) scored the AI-generated feedback from Claude Sonnet 4 using 12 systematically developed prompts to elicit feedback. Raters were asked to score feedback produced by these prompts across four areas of evaluation: pedagogy, linguistics, Islamic content, and AI reliability. The highest rated configuration was Prompt 4 (Feedback-only sequencing, English Lecturer Persona), with a mean rating of 31.00 (out of 35) in all categories. A Friedman test showed there were statistically significant differences among the four evaluative categories, $\chi^2(3) = 30.077$, $p < .001$. Additionally, inter-rater reliabilities were high for each of the possible pairs of raters ($r = .89$ -.96). Overall, this research suggests that prompt engineering is a potentially viable method of pedagogical mediation, allowing educators to develop more culturally responsive and pedagogically relevant AI-generated feedback systems for Islamic EFL higher education settings.**

Keywords*—AI-generated feedback; pedagogical mediation; prompt engineering

I. Introduction

AI-generated feedback has transformed the landscape of educational technology and second-language writing instruction. The new generation of large language models has allowed language models to create long, well-written, and contextually sensitive text-based feedback, much like human feedback in terms of content and purpose [1], [2]. This has led to an increase in academic interest in using AI-generating feedback to assist students in developing the skills required to write in the English as a foreign language (EFL) context [3], [4], such as accuracy in language use, argumentative coherence, and academic discourse [5], [6].

On the other hand, feedback is seen as the primary means of helping students to evaluate and revise their writing [7]. Traditional methods of giving feedback are, however, hampered by the practical realities of teaching at scale and include large class sizes, short periods of contact between teachers and students, and varying amounts of teacher workloads, leading to either delayed or little feedback [8], [9]. Consequently, researchers have turned to AI-generated feedback as an additional resource for teachers to use when teaching writing to provide immediate and personalized feedback to students [10], [11]. Studies that have examined learners' views of AI-assisted feedback have generally found them to be supportive of using AI-assisted feedback, citing the potential for improved revision processes [12], [13]. Concerns remain about the pedagogic value, cultural relevance, and epistemological validity of such feedback [14], [15].

One of the most important areas that has been studied less frequently than others when evaluating the quality of AI-created feedback is prompt engineering. Prompt engineering means deliberately designing the instructions provided to a language model for the purpose of generating the desired output [16], [17]. There is growing evidence that how a prompt is designed can influence the nature, quality, and utility of the model's output [18], [19]. Furthermore, many applied linguistics studies have treated prompts as neutral technical input to an educational technology rather than as reflective of the educator's intention for instruction or the epistemological structure of their knowledge system [20], [21]; this treatment of prompts has obscured the contribution that the prompt makes toward the quality of the generated feedback.

This issue is further complicated in educational settings that involve student writing based on culturally and religiously informed knowledge systems. For example, in Islamic higher education settings, EFL argumentative writing typically includes references to the Qur'an, Hadith, and Islamic ethics within academic arguments, making the need for feedback that goes beyond linguistic correctness [22], [23]. Thus, the same piece of written feedback needs to meet multiple criteria, such as, to be pedagogically constructive; linguistically suitable for the learner's level of proficiency; factually accurate and respectful in its treatment of Islamic knowledge; and reliable in avoiding factual inaccuracies or epistemological overstepping [24], [25]. Generic AI prompts without contextual and epistemological constraints may generate feedback that is at odds with the intended educational objectives or religious sensitivities, creating both ethical and pedagogical dilemmas [26], [27].

Those problems would be addressed by examining how different prompt engineering approaches affect expert evaluations of LLM-generated feedback on EFL argumentative essays that incorporate Islamic perspectives. Unlike previous studies that examine learner outcomes resulting from AI-assisted feedback, expert-based evaluation framework was used that assumes feedback must first receive pedagogical and epistemological validation before being implemented in the classroom [28], [29]. By intentionally manipulating the

*Corresponding author.

prompting patterns and task instruction sequences, prompt designs that generate feedback would be determined, that experts judge to be pedagogically effective, linguistically accurate, religiously appropriate, and technologically reliable.

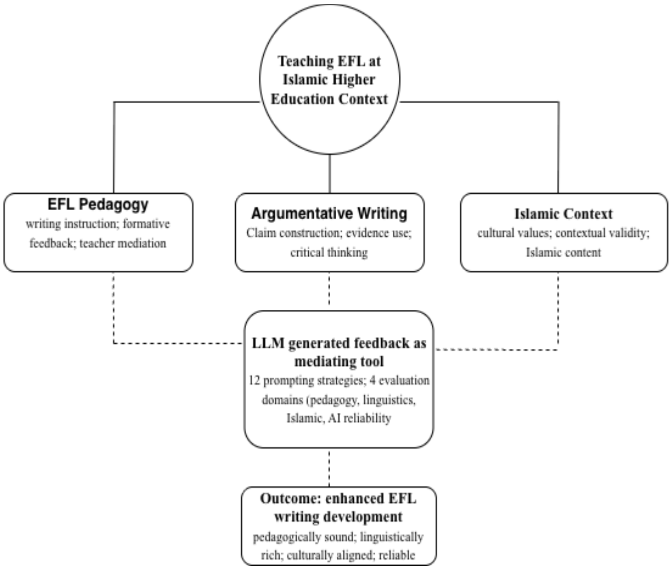
Teaching EFL at Islamic Higher Education Context
EFL Pedagogy
writing instruction; formative feedback; teacher mediation
Argumentative Writing
Claim construction; evidence use; critical thinking
Islamic Context
cultural values; contextual validity; Islamic content
LLM generated feedback as mediating tool
12 prompting strategies; 4 evaluation domains (pedagogy, linguistics, Islamic, AI reliability
Outcome: enhanced EFL writing development
pedagogically sound; linguistically rich; culturally aligned; reliable

Fig. 1. Conceptual framework of the study.

The major goal of the current study is to examine how Prompt Engineering affects the Quality of Feedback from AI in Islamic English as a Foreign Language (EFL) writing. In addition to defining prompts as scaffolds that affect how LLMs reason as pedagogically mediated, mediation also includes two additional categories: the structure of personae determines an instructional tone/ authority and the sequential order of tasks determines whether formative or evaluative purposes are being followed. This will be measured using ratings by experts based on four areas of pedagogy (how actionable it is), linguistics (its accuracy), religious content (it is sensitive to Islam), and reliability (the likelihood of hallucination), with a total of 28 measurable criteria. The second purpose is to explore the use of prompt engineering in an Islamic context, including the need for both Academic and Religious validation. The third purpose is to identify a scalable solution to reduce teacher workload in large classes. Fig. 1 presents the conceptual framework of the study.

This study investigated three research questions: 1) To what degree do two types of prompt engineering methods (use of a persona-based or base prompt; use of different task sequences) affect the quality of AI-generated written feedback in terms of pedagogy, linguistics, Islamic content, and reliability? 2) Which of the several possible prompt configurations will generate the most highly rated written feedback by experts for an EFL student's argumentative writing that includes Islamic themes? 3) How reliable are expert ratings of the quality of written feedback and how can rationales account for differences in the quality of the written feedback?

II. Literature Review

Most early automated writing evaluation systems evaluated EFL student writing based on surface-level linguistic characteristics. It made those systems were criticized for failing to adequately evaluate higher-order concerns, such as the organization and rhetorical effectiveness of students' writing, which are essential to the development of students' academic writing [7]. The advent of large language models then has dramatically increased the potential for AI-generated feedback to assess a wide range of EFL writing concerns, including those related to higher-order concerns, and to generate a wider variety of responses [30], [31].

Studies that have recently investigated LLM-generated feedback have found that the feedback can assist EFL learners in their revision processes, increase their engagement with writing assignments, and enhance learner autonomy in EFL contexts [32], [33]. Studies have also indicated that learners often view AI-generated feedback as immediate, non-judgmental, and accessible, and therefore likely to decrease affective barriers that may exist between learners and traditional teacher feedback [34], [35]. Nevertheless, the same studies have also demonstrated considerable variation in the quality of AI-generated feedback, with some LLM-generated feedback being vague, overgeneralized, or providing feedback that is inconsistent with instructional goals and genre expectations [36], [37].

Given this mixed evidence, it appears that the effectiveness of AI-generated feedback is contingent upon how the feedback system is designed and implemented in the classroom [7]. Researchers have increasingly emphasized that AI feedback systems need to be examined not only in terms of their technical capability, but also in the broader instructional context in which they will be used, including the nature of the tasks, the type of feedback, and the context in which the learner is located [13], [38].

Prompt engineering has become a key aspect in determining the behavior and the quality of output of large language models and is viewed as strategies to optimize performance or improve output accuracy by specifying words and/or context carefully [18], [39]. From an educational perspective, however, prompts represent instructional design choices that specify the epistemic position of the model such as what does the model know, the communicative role and the pedagogical priorities [18].

It has been shown in the technical literature that prompts that specify a persona influence the tone, level of explanation and reasoning pattern exhibited by the model's generated responses [40], [41]. Models that are instructed to take a specific persona tend to generate more structured, supportive and pedagogically-oriented feedback compared to models that are prompted neutrally or minimally [42], [43]. It has also been demonstrated that the sequence in which tasks are presented to a model can significantly influence the generated response. For example, requests that first emphasize scoring/evaluation may inhibit elaboration and shift the model towards justifying a score rather than providing support for the writer [43], [44].

Although researchers have identified various influences of prompt engineering in technical literature, the application of prompt engineering has received very little attention in applied linguistics research, with most studies relying on single or ad hoc designs for prompts, thus limiting the ability to separate the effect of prompt design from the inherent abilities of the model [16], [26]. Therefore, there is a significant need for empirical studies to examine prompt design as an instructional variable and explore its effect on feedback quality across multiple dimensions, such as pedagogy, language, and contextualization [18], [19]. Assessment-for-learning approaches emphasize the formative function of feedback, and specifically argue that feedback should be guided by actionability, developmental orientation, and learner-centeredness [45], [46].

Task sequencing in AI-generated feedback has direct implications for these principles. If a model is required to assign a score before generating feedback, then the model is likely to provide feedback that is evaluatively based rather than pedagogically based, and therefore provides feedback that explains why it made a particular grade rather than how the writer can learn from their mistakes [6], [47]. On the other hand, if a model is requested to generate feedback before assigning a score, then the model is likely to produce more diagnostic and formative feedback, and thus feedback that is more aligned with assessment-for-learning principles [48], [49]. Although this assumption is theoretically supported, empirical evidence regarding whether task sequencing impacts the quality of AI-generated feedback remains minimal, and, therefore, there exists a need for systematic research [12], [50].

In Islamic higher education contexts, EFL writing is not solely a linguistic exercise, but rather a tool that allows students to articulate religious and ethical viewpoints within academic discourse [48], [49]. Students may incorporate Quranic verses, Hadith and Islamic moral reasoning into argumentative essays in order to justify their claims, and therefore, teachers must provide feedback that is sensitive to both academic norms and religious epistemology [50], [51].

In culturally responsive teaching, teachers are expected to provide students with constructive, academic feedback which supports student learning by respecting their own cultural and epistemological views [52], [53]. This means, for example, that when providing AI-generated feedback on an Islamic-integrated written work, teachers need to have knowledge of what Islamic-religious texts were used as reference and be able to communicate them respectfully; and also support learners to combine academic reasoning with faith-based reasoning [52], [53]. As a result, if an educator provides learners with AI-generated feedback which does not demonstrate the same level of respect as a teacher would, the AI-generated feedback may not accurately reflect the religious content, will simplify the learner's theological views and thus, diminish or undermine the learner's identity as an epistemic being. Therefore, the use of AI-generated feedback is likely to raise significant pedagogic and ethical concerns [23], [54].

Research on AI and religion has also highlighted the concern that language models may generate inaccurate or misleading religious content without proper constraint [55], [56]. These concerns illustrate the importance of designing prompts in a manner that guides the behavior of AI systems in sensitive areas. However, research on how prompt engineering influences the accuracy of religious references and sensitivity of AI-generated feedback in educational contexts has yet to be explored empirically, especially in the context of EFL writing research [57], [58].

As noted previously, the literature reviewed above demonstrates several gaps in knowledge. First, although numerous studies have examined AI-generated feedback in EFL writing, relatively few studies have investigated how different prompt engineering techniques affect feedback quality across pedagogical, linguistic and ethically-related dimensions [18], [19]. Second, research on using personas in prompting and task sequencing has generally remained theoretical or anecdotal in the field of Applied Linguistics, with few studies having empirically validated the use of either technique through expert evaluations [15], [59]. Finally, the context of Islamic-integrated EFL writing has yet to appear in empirical studies on the use of AI-assisted feedback, despite the pedagogical and ethical relevance of this area of study [60].

Therefore, a systematic expert evaluation of LLM-generated feedback of twelve different prompt design parameters was conducted. Unlike other studies that have tested 1 to 5 ad hoc prompts for a single domain, the current study manipulated two theoretically grounded parameters (persona × sequence = 12 conditions), with culturally sensitive evaluations of multiple dimensions. The comparative analysis of existing prompting engineering research with this study is presented in Table I.

TABLE I. COMPARATIVE ANALYSIS OF PROMPT ENGINEERING RESEARCH.

Study	N Prompts	Domains Tested	Islamic Context	Stats
Gozzi & Di Maio (2024)	2	General	No	Wilcoxon, McNemar, Friedman Test
Choi et al (2025)	2	Pedagogy	No	ANOVA
Chen (2025)	2	General	No	Descriptive
This study	12	4 domains	Yes	ANOVA, Kruskall Wallis, Friedman, Wilcoxon Signed-Rank

The theoretical voids that this study has addressed include (prompts as pedagogical tools and not neutral inputs), an Islamic EFL environment that is unexplored and a systematic 2 × 3 × 2 design that will be validated statistically.

III. Methodology

A quantitative expert evaluation methodology is used to investigate how various prompt engineering strategies affect the quality of large language model (LLM) generated feedback on EFL students' argumentative essays that include Islamic perspectives. Using expert evaluation was appropriate due to the fact that establishing pedagogically valid, epistemologically accurate, and ethically responsible AI in educational environments requires the establishment of validity before it is implemented into classroom instructional practices, particularly so in religiously grounded content areas, where inappropriate or incorrect feedback may cause repercussions that go beyond developing language skills to issues of epistemological trust and ethical responsibility.

The materials used in this study are the ninety genuine B1-B2 argumentative essays from Indonesian Islamic university EFL students, which included Quranic verses/Hadith, and were all examples of routine coursework as is typical for many other students (not revised for ecological validity) in this study. The incorporation of Islamic perspectives included references to Quranic verses, Hadith, Islamic moral principles, and religiously-grounded reasons to support claims and counter-claims, thus requiring feedback that addressed both the academic and religious aspects of writing [60].

The main independent variable examined was prompt engineering strategy. Twelve discrete prompt strategies were developed through a systematic combination of two prompting parameters: prompting pattern and task instruction sequence. Table II presents the prompting pattern and task instruction type implemented in this study. Viewing prompt design as an independent instructional parameter aligns with contemporary recommendations in applied linguistics to explore how AI-generated output is impacted by human-designed instructional inputs as opposed to attributing results to model characteristics alone [61].

TABLE II. PROMPTING PATTERN AND TASK INSTRUCTION TYPE

No	Prompting Pattern	Task Instruction Type
1	Base	Feedback only
2	Base	Scoring to Feedback
3	Base	Feedback to Scoring
4	English Lecturer	Feedback only
5	English Lecturer	Scoring to Feedback
6	English Lecturer	Feedback to Scoring
7	English Lecturer in Islamic Higher Education	Feedback only
8	English Lecturer in Islamic Higher Education	Scoring to Feedback
9	English Lecturer in Islamic Higher Education	Feedback to Scoring
10	Islamic Cleric	Feedback only
11	Islamic Cleric	Scoring to Feedback
12	Islamic Cleric	Feedback to Scoring

The prompting pattern parameter was differentiated between base prompts and persona-based prompts. Base prompts supplied task instructions that assigned no specific role or contextual identity to the language model, representing typical ad hoc uses of AI in educational contexts. By contrast, persona-based prompts directed the model to assume a particular professional or epistemic role, a factor previously demonstrated to influence the degree of explanatory detail and pedagogical orientation in LLM-generated output [61]. Three personas were used: English lecturer, English lecturer in Islamic higher education, and Islamic cleric. These personas were selected to represent varying degrees of pedagogical, disciplinary, and religious authority, which are all relevant to the multiple forms of epistemic authority that are used in evaluating EFL writing that incorporates Islamic perspectives. Prior research has indicated that specifying roles will influence the models to reason within domain-specific patterns and styles of communication, which indicates that selecting a persona is a theoretically justified design parameter and not simply a superficial variation.

While the original experimental design involved writing essays that were never in conflict with one another, the evaluation system of all of the prompts was uniform by virtue of the use of the same rubric to score (Pedagogy, Linguistics, Islamic Content, and AI Reliability); therefore, comparisons could be made on the evaluation criteria level. Each prompt was rated against the same four dimensions as other prompts, thus enabling both an analysis of consistency within a prompt and a cross-prompt comparison based upon aggregated ratings. Therefore, while this does not represent a completely controlled repeated-measures design, it does provide a structured framework for conducting comparative statistical analyses. All the prompts were structured in a similar way. The prompts contained: 1) background to the essay (90 argumentative essays for intermediate EFL students using references from Islam); 2) a character if one was included; 3) how to complete the task; 4) 4-domain assessment criteria; 5) how to present the work you did as an outcome of the task (approximately 150–200 words). The example of a prompt is presented in Table III.

TABLE III. REPRESENTATIVE PROMPT EXAMPLE

Prompt	Type	Example Structure
P1	Base, feedback only	Provide feedback on the EFL essay. No score
P4	Lecturer, feedback only	You are an English lecturer at an Islamic University. Assess the essay using the given rubric and provide the feedback. No score needed.
P12	Cleric, feedback to scoring	You are cleric. Assess the essay using the given rubric. Provide feedback first then score.

The second parameter concerned the task instruction sequence. Three sequences were used. In the feedback-only sequence, the model was directed to provide feedback but was not asked to assign a numerical grade. In the scoring-to-feedback sequence, the model was directed to assign a grade prior to providing explanatory feedback. In the feedback-to-scoring sequence, the model was directed to provide detailed explanatory feedback prior to assigning a numerical grade based upon the feedback provided. The distinction between these three sequences represents the theoretical foundation for assessment-for-learning theory, which advocates for formative guidance to occur prior to summative judgment in order to optimize instructional effectiveness [7]. The combination of the prompting pattern and the task sequence resulted in twelve unique prompt strategies, each designed to elicit a particular set of reasoning, tone, and instructional foci.

To use a balanced incomplete block design as design validity, every prompt was tested in seven to eight essays (which provides for ninety per cent total coverage), with each essay having one or two prompts at most (a way to help control fatigue). Experts coding/blinding to prompt source rated each of the four Claude Sonnets (the lowest level of Islamic hallucinations per pilot compared to ChatGPT/Gemini).

A total of eight specialists took part in the assessment procedure. Four first-rate assessors and four second-rate assessors came from 5 Islamic universities in Indonesia. The experts are the specialists from 4 fields, namely English Education, Linguistics, Islamic Studies, and Educational Technology programs. The inclusion of assessors from various universities improved the generalizability and transferability of the findings by limiting the potential effects of institutional norms or practices on the results. All the assessors either held PhD's or were professional practitioners who had considerable experience in their own areas; this fulfills the minimum requirements for designating an individual as an expert in educational research.

Second-rate assessors were added to the study to allow for assessing the reliability and consistency of the evaluations provided by the primary experts. The problem of subjectivity in expert-based assessment procedures can be addressed by using second-rate assessors. The second-rate assessors used the same evaluation guidelines and instructions as did the first-rate assessors and independently assessed the same sets of sample feedback, allowing for statistical assessments of the inter-rater reliability and increasing the overall methodological rigor of the study.

The pedagogy mediation definition refers to the implementation of prompt strategies (persona as structure; sequence as format) as instructional scaffolds shaping LLM reasoning. The generated feedback was assessed through 4 domains and 28 criteria. Pedagogical quality was assessed in terms of organization, coherence, actionable suggestions, constructive suggestions, tone, cognitive load management, and developmental progression, consistent with theories of formative feedback [62]. Linguistic quality was assessed in terms of accuracy, accessibility, appropriateness to learner proficiency level, clarity of explanation, and priority given to language issues, consistent with central themes of teaching EFL writing [63]. Islamic content quality was assessed in terms of accuracy of Quranic and Hadith references, accuracy of Islamic concepts and terminology, cultural and religious sensitivity, integration of Islamic evidence, and balance of perspectives, consistent with principles of culturally responsive pedagogy [60]. AI naturalness was assessed in terms of human likeness, contextual relevance, factual accuracy, logical consistency, and avoidance of hallucinations, consistent with emerging concerns in AI ethics and reliability [15].

Each criterion was rated on a five-point scale, with higher ratings indicative of higher quality. Domain ratings were determined by summing the ratings of the individual criteria for each domain, while overall ratings for each prompt strategy were determined by averaging the ratings of the four domains. This rating scheme permitted both domain-by-domain comparisons of the prompt strategies and overall comparisons of the prompt strategies, consistent with prior studies examining expert evaluations of computer-assisted language learning. This process ensured that all expert opinions were truly independent in their evaluation of the prompt strategies; as well as non-comparative opinions based on multiple exposures to the same text.

The analysis was conducted through first, calculation of mean ratings for each prompt strategy across the four domains [58]. Second, Kruskal-Wallis, Friedman and Wilcoxon Signed-Rank Test were conducted as statistical validation and third the Pearson product-moment correlation coefficients were calculated for each domain to determine the extent of agreement between the ratings of the primary experts and secondary raters.

IV. Results

Expert assessments clearly demonstrated considerable and consistent differences in quality of feedback across the 12 strategies examined, across all 4 assessment domains (pedagogy, language, religion, AI).

Pedagogical quality results for all twelve prompt strategies are provided in Table IV. There were seven criteria used to determine pedagogical quality, including organization (1), coherence & flow (2), comprehensiveness (3), actionability (4), constructiveness & tone (5), cognitive load management (6), and developmental progression (7); each criterion was evaluated on a scale of 0-5. The result of pedagogical evaluation is presented in Table IV.

TABLE IV. Pedagogical Evaluation Result

Prompt	Pedagogical Approach							Total
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
P4	5	5	5	5	5	5	5	35
P5	3	5	5	5	5	5	5	33
P7	4	5	4	5	5	5	5	33
P11	5	5	3	4	5	5	5	32
P6	3	5	4	5	5	4	4	30
P1	4	4	5	3	5	3	5	29
P8	4	4	3	5	3	5	5	29
P10	5	5	3	5	3	3	4	28
P12	3	4	5	4	3	5	3	27
P3	2	4	3	5	5	5	3	27
P9	4	2	3	4	4	5	3	25
P2	4	4	3	2	5	4	2	24

Prompt 4 (English lecturer persona with feedback-only task type) achieved the highest pedagogical score (35/35), with a maximum score across all seven criteria. The expert evaluation scores indicated that the combination of a pedagogically grounded persona and a purely formative task sequence allowed the AI to emulate teacher-like practices in providing feedback most effectively.

Prompt 5 (feedback-only task instruction type, combined with English lecturer persona) and prompt 7 (feedback-only task instruction type, combined with English lecturer in Islamic higher education persona) both obtained total pedagogical scores of 33. Prompt 11 (Islamic cleric persona with a scoring-to-feedback task instruction type) achieved a total pedagogical score of 32.

Moderate pedagogical performance was observed for prompts 6, 1, and 8, which received total scores ranging from 29 to 30. Prompt 6 used a feedback-to-scoring sequence and the English lecturer persona, prompt 8 used a scoring-to-feedback sequence and the English lecturer in Islamic higher education persona, and prompt 1 used a base prompt with a feedback-only instruction type.

Lower pedagogical effectiveness was observed for prompts 10, 12, and 3, which involved either the Islamic cleric persona with feedback-only or feedback-to-scoring sequences or the base prompt with feedback-to-scoring instruction. Total scores for these prompts ranged from 27 to 28, reflecting limitations in cognitive load management and developmental sequencing. The lowest pedagogical scores were recorded for prompts 9 and 2, which corresponded to the English lecturer in Islamic higher education persona with feedback-to-scoring and the base prompt with scoring-to-feedback, respectively. These prompts received notably low ratings in coherence, actionability, and developmental progression, resulting in total scores of 25 and 24.

The evaluations of the linguistic quality of the AI-generated feedback based on the expert evaluation are shown in Table V.

TABLE V. LINGUISTIC EVALUATION RESULTS

Prompt	Linguistic Aspect							Total
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
P6	5	5	5	5	5	5	5	35
P10	5	5	5	5	4	5	5	34
P3	5	5	5	5	4	5	5	34
P4	5	5	5	5	5	4	5	34
P11	5	5	5	4	4	5	5	33
P5	5	5	5	4	4	5	5	33
P7	4	5	5	5	5	5	4	33
P8	4	5	5	5	5	5	4	33
P9	4	5	5	5	5	5	4	33
P2	4	5	5	5	4	5	4	32
P1	4	5	5	4	4	5	4	31
P12	4	5	5	4	4	5	4	31

Table V shows the maximum linguistic score, which is 35 points (the sum of 7 linguistic criteria) and these include; vocabulary and grammar feedback accuracy (1), mechanics and conventions (2), linguistic accessibility of feedback (3), proficiency level appropriateness (4), clarity of explanations (5), prioritizing language issues (6), and overall pedagogical effectiveness of Language Feedback (7).

Prompt 6 had the best linguistic score (35/35) and combined the English Lecturer persona with a feedback-to-scoring task instruction sequence. The experts gave this combination a perfect score in all of the linguistic criteria and therefore concluded that the feedback produced with this prompt was perfectly clear in its linguistic content and was appropriate in relation to student's proficiency levels.

Both Prompts 10 and 3 achieved linguistic scores of 34 and although the two prompts differ greatly in the prompting configurations, the results for both show good consistency in vocabulary and grammar accuracy, mechanical and accessibility aspects. Lower scores for the clarity of explanations suggest that, although the AI produced linguistically correct responses, the instructional explicitness of the responses was marginally less consistent than in the case of Prompt 6. Prompt 4, which combined the English Lecturer persona with a feedback-only task sequence, also achieved a high linguistic score of 34.

Both Prompts 11 and 5, both using the scoring-to-feedback sequencing with the Islamic Cleric persona (Prompt 11) and the English Lecturer persona (Prompt 5), respectively, achieved linguistic scores of 33. These results indicated generally strong linguistic accuracy and appropriateness but revealed some minor reductions in the clarity of explanation and proficiency-level calibration. Prompts 7, 8, and 9, all using the same persona (English Lecturer in Islamic Higher Education) and different task instruction types, each achieved linguistic scores of 33.

Moderate linguistic performance was noted for Prompts 2, 1, and 12, with total scores ranging from 31 to 32. Prompt 2 combined the base prompting pattern with the scoring-to-feedback, Prompt 1 used the base prompt with the feedback-only, and Prompt 12 employed the Islamic Cleric persona with the feedback-to-scoring. These prompts maintained satisfactory levels of linguistic accuracy and accessibility but were hampered by lower ratings in clarity of explanation and prioritization of language issues.

Overall, the results of the linguistic assessments indicate that high-quality language feedback can be produced through various combinations of prompting and task instructions, but the highest levels of linguistic effectiveness were found when a pedagogically relevant persona was combined with a task instruction sequence that permits a detailed explanation before evaluation.

Table VI contains expert evaluations on the accuracy and relevance of Islamic content based on twelve different methods of promoting Islamic content.

A maximum of 35 points could have been scored for the Islamic quality of the content; seven criteria were evaluated: the accuracy of references to Qur'anic and Hadith (1), the Islamic concepts and terminology used (2), the respectfulness and tone used (3), the cultural suitability (4), the integration of Islamic evidence (5), the balance of perspectives (6), and the overall accuracy of religious referencing (7). Scores of 31 were achieved for Islamic content by two prompts: Prompts 4 and 8. Each prompt had a total score of 31, although they differed in their configuration. Prompt 4 had the English lecturer persona and a "feedback-only" instructional task, whereas Prompt 8 had the English lecturer in Islamic Higher Education persona and an instructional task using a "scoring-to-feedback" sequence.

These results demonstrate that locating the AI within an educational framework, whether general or institutional Islamic, supports the AI's ability to produce responses that are both accurate and respectful of Islamic sources, especially when the instructional tasks are designed to promote structured feedback. Each of Prompts 2, 5, 6, and 7 received a total score for Islamic content of 30, even though they differed in terms of their configuration of promotional patterns. Prompt 2 was configured using the base promotional pattern and a "scoring-to-feedback" sequence; Prompt 5 was configured using the English lecturer persona and a "scoring-to-feedback" sequence; Prompt 6 was configured using the English lecturer persona and a "feedback-to-scoring" sequence; and Prompt 7 was configured using the English lecturer in Islamic Higher Education persona and a "feedback-only" sequence.

TABLE VI. ISLAMIC CONTENT EVALUATION RESULTS

Prompt	Islamic Content							Total
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
P4	5	4	4	4	4	5	5	31
P8	5	4	4	5	4	4	5	31
P2	5	4	4	4	4	4	5	30
P5	5	4	4	4	4	4	5	30
P6	5	4	4	4	4	4	5	30
P7	5	4	4	4	4	4	5	30
P11	4	4	4	4	4	4	4	28
P12	4	4	4	4	4	4	4	28
P3	4	4	4	4	4	4	4	28
P9	4	4	4	4	4	4	4	28
P15	4	4	4	4	4	4	4	28
P10	4	4	4	5	3	3	4	27

Moderate Islamic content performance was reported for Prompts 11, 12, 3, and 9; each of these prompts achieved a total score of 28. These prompts included several variations of Islamic clerics as personas (Prompts 11 and 12), base prompts (Prompt 3), and the English lecturer in Islamic Higher Education persona and a "feedback-to-scoring" sequence (Prompt 9).

Prompt 10 was found to be the least effective in terms of Islamic content, with a total score of 27; it utilized the Islamic cleric persona and a "feedback-only" task sequence. Although this prompt received high ratings for cultural appropriateness and respectfulness, its performance was low for Islamic evidence integration and balance of perspectives. The lowest Islamic content score was reported for Prompt 1; this prompt corresponded to the base promotional pattern and "feedback-only" instructional task; it achieved a total score of 24. As such, Prompt 1 received the lowest ratings for Quranic and Hadith reference accuracy and integration quality.

In summary, the results related to Islamic content clearly demonstrate that religious accuracy and sensitivity in AI-generated feedback depend on the interactions between the promotional patterns and task instructional types. Prompt strategies that utilize pedagogical personas and structured task sequences, especially those framed with educational rather than purely religious authority, are most likely to achieve higher evaluations from experts.

Table VII shows the expert evaluation results for AI naturalness and reliability on twelve different prompt strategies. The maximum possible score for this domain is 25, based upon five criteria: language naturalness (1), contextual responsiveness (2), student text accuracy (3), factual knowledge accuracy (4), and logical consistency (5). This domain evaluated how closely an AI's generated feedback mimics human-written responses in terms of content while being factually and logically reliable.

TABLE VII. AI RELIABILITY EVALUATION RESULTS

Prompt	AI Reliability					Total
	(1)	(2)	(3)	(4)	(5)	
P4	5	5	4	5	5	24
P2	4	4	5	4	5	22
P3	4	5	5	3	3	20
P6	3	4	4	5	4	20
P7	4	5	3	3	5	20
P9	4	3	4	5	4	20
P10	4	3	4	4	3	18
P12	4	4	3	4	3	18
P1	3	3	4	4	3	17
P5	2	3	5	4	3	17
P8	3	3	2	4	4	16
P11	4	3	3	2	3	15

Prompt 4 received the highest naturalness & reliability score (24/25). The evaluators assigned this configuration the maximal rating for language naturalness, contextual responsiveness, factual knowledge accuracy, and logical consistency, but slightly lower than the maximal rating for student text accuracy. Prompt 2 had the second-highest score (22), although it employed the base prompting pattern with a scoring-to-feedback sequencing. However, the lower scores in language naturalness and contextual responsiveness indicate that the absence of a persona-based frame constrained the perceived human-likeness of the generated feedback.

Moderate naturalness and reliability performance was observed among prompts 3, 6, 7, and 9; each prompt achieved total scores of 20. These four prompts represented different configurations including base prompting with a feedback-to-scoring (prompt 3), English lecturer persona with feedback-to-scoring (prompt 6), English lecturer in Islamic higher education persona with feedback-only (prompt 7), and English lecturer in Islamic higher education persona with feedback-to-scoring (prompt 9). The four prompts were given uniformly high ratings by evaluators, although they had some areas where there was room for improvement (logical consistency, and/or student text accuracy).

The lowest naturalness and reliability scores were recorded for prompts 10 and 12, both which had total scores of 18. Prompt 10 employed the Islamic cleric persona with feedback-only instructions, while prompt 12 combined the Islamic cleric persona with feedback-to-scoring. Although the two prompts demonstrated acceptable levels of language naturalness, the evaluators identified reduced performance in terms of logical consistency and hallucination avoidance. The lowest naturalness and reliability scores were observed for prompts 1, 5, 8, and 11, with total scores ranging from 15 to 17. These four prompts included base prompting with feedback-only (prompt 1), English lecturer persona with scoring-to-feedback (prompt 5), English lecturer in Islamic higher education persona with scoring-to-feedback (prompt 8), and Islamic cleric persona with scoring-to-feedback (prompt 11).

An aggregation of the expert evaluations for each of the five prompt strategies on all four domains of AI-generated feedback quality is presented in Table VIII.

TABLE VIII. OVERALL EVALUATION RESULTS

Prompt	Overall Aspects				Total Score
	Pedagogy	Linguistics	Islamic Content	AI Reliability	
P1	29	31	24	17	101
P2	24	32	30	22	108
P3	27	34	28	20	109
P4	35	34	31	24	**124**
P5	33	33	30	17	113
P6	30	35	30	20	115
P7	33	33	30	20	116
P8	29	33	31	16	109
P9	25	33	28	20	106
P10	28	34	27	18	107
P11	32	33	28	15	108
P12	27	31	28	18	104

Based upon the average scores for the four domains of quality, Prompt 4 received the highest average score of 124, exceeding the average scores for all other prompt strategies based upon the aggregated assessment of quality across all four domains. The English Lecturer persona was utilized in Prompt 4, with a "feedback-only" task instruction sequence, and received among the highest ratings for each of the four domains of quality; the highest rating for Pedagogical Quality (35), the second-highest rating for Linguistic Quality (34), the highest rating for Islamic Content Accuracy (31) and the highest rating for AI Naturalness/Reliability (24).

Descriptive statistics were used to determine performance averages on each of the test criteria. Table IX presents performance across evaluation aspects.

TABLE IX. PERFORMANCE ACROSS EVALUATION ASPECTS

Aspects	Mean	SD
Pedagogy	29.3	3.3
Linguistics	33.0	1.4
Islamic Content	28.8	2.2
AI Reliability	18.9	2.7

The linguistics area had the best average score with the least amount of variation from prompt to prompt; thus, it is clear that the responses given in this area are consistent across all prompts. Conversely, the AI Reliability area had the worst average and the most variation among prompts, suggesting weak and unstable performance. The Pedagogy and Islamic Content areas both had a medium range for their averages and variation.

Since this was an exploratory study using a small sample and since the evaluations by experts are ordinal data, both parametric and non-parametric statistical tests were used. The first objective of this exploratory study is to determine if the strategies for prompting result in different performance.

First, a one-way ANOVA test was conducted in order to determine if there were significant differences among the means of the twelve different prompting strategies. The results were F (11, 36) = 0.234, p = .993, which indicated that there were no statistically significant differences among the prompts.

A Kruskal-Wallis test was then run to determine if there were differences in the total scores of the students. The result is presented in Table X.

TABLE X. KRUSKAL-WALLIS TEST

Test	Value
Kruskall-Wallis H	16.92
df	11
Asymp.Sig	0.109

While variations in the total scores were found among the students, no statistical evidence was found that the differences in the total scores could be attributed to the differences in the design of the prompting strategies.

A Friedman test was also conducted to compare the students' scores from the four evaluation criteria.

TABLE XI. FRIEDMAN TEST

Test	Value
N	12
Chi square	28.74
df	3
Asymp.Sig	0.000003

Based on Table XI, statistical evidence indicated that the differences in the scores of the students across the four evaluation aspects were statistically significant.

Following the Friedman test, post hoc Wilcoxon signed-rank tests with Bonferroni correction were also performed. The results of those tests indicate that:

TABLE XII. WILCOXON SIGNED-RANK TEST

Test	Value
AI Reliability vs. Pedagogy	p ≈ 0.002
AI Reliability vs. Linguistics	0.0005
AI Reliability vs. Islamic Content	0.001

Based on the Wilcoxon signed-rank test provided in Table XII, AI Reliability scores are significantly lower than all other criteria.

An effect size analysis was also conducted to determine the magnitude of the differences. The result is presented in Table XIII.

TABLE XIII. EFFECT SIZE ANALYSIS

Test	Value	Effect
Kruskal–Wallis ε^2	0.12	moderate effect
Friedman Kendall's W	0.80	strong effect

Based on Table XIII, while the prompting strategies do not have a statistically significant impact on the total scores of the students, the differences in the scores across the evaluation aspects are large and consistent.

In addition, the Pearson correlation coefficient test was also conducted, with the result between primary expert evaluations and secondary inter-rater evaluations across all four domains ranged from .89 to .96, indicating an extremely high level of inter-rater agreement.

Based on the statistics above, it can be inferred that although Prompt P4 had the top overall score, the lack of statistical significance (p = 0.109) indicates that the differences in scores should be interpreted as descriptive trends and not necessarily as evidence that Prompt P4 is superior to other prompts. It is interesting to note that the prompts that are higher ranked are generally more balanced across the four evaluation criteria, rather than being high-scoring in only one area.

V. DISCUSSION

The association of the mediation of teaching through prompt engineering with enhancements in the quality of AI-generated student feedback for all aspects of evaluation is evident. There is no indication in the findings that there is a direct cause-and-effect relationship. Instead, thoughtful prompt design may enable the production of feedback that is more like teacher-like reasoning. It is possible that the instructional framing and domain-specific knowledge that the LLM received through the persona prompts used in the study enabled the LLM to generate responses that are structurally and functionally similar to those generated by teachers; however, the specific mechanisms through which this similarity occurs will need to be explored further.

The observation of feedback first in the sequence of the higher performing prompts is consistent with the assessment-for-learning model. In the assessment-for-learning model, the purpose of the feedback is to assist learners in improving their work rather than in assessing what they have accomplished. Although there is no basis in the current data to conclude that the sequencing of feedback, as described above, is a causal factor in improving learning outcomes, the expert ratings suggest that the sequencing may favor the formative purposes of feedback over the evaluative purposes of feedback in ways that are consistent with established principles of writing pedagogy. The P4 configuration appears to be the highest rated configuration due to the fact that the lecturer persona provides sufficient pedagogical framing and does not extend into the area of clerical or administrative tasks, and the fact that the feedback-only sequencing appears to optimize the potential for the scaffolding of the generated output. As with all interpretations of this sort, they are tentative until replicated in a classroom setting.

The results suggest that the quality of AI-generated feedback for argumentative writing in EFL is not a function of the intrinsic capabilities of the LLM alone, but is also substantially influenced by the design of the prompting strategies. The variation that was observed across the Pedagogical, Linguistic, Islamic Content, and AI Reliability domains is consistent with a growing body of research that indicates that the effectiveness of AI-assisted feedback is contingent upon the design decisions that inform the LLM of the instructional objectives and needs of the learner [19], [64]. Therefore, the current study contributes to comparative research on AI and teacher feedback by providing evidence that AI feedback should not be considered as a singular, stable entity, but rather as a variable output that is likely influenced by how prompt design influences the LLM's interpretation of instructional objectives [6], [65].

A major finding of this study is that prompts that assigned the role of an English Lecturer were associated with the highest average ratings for pedagogy and overall averages when compared to the other prompt types. Expert and inter-rater rationales indicated that the instructional sequencing was more apparent, the scaffolding was more explicit, and the alignment with the principles of learner-centered feedback was more apparent for the prompts that assigned the role of an English Lecturer. These interpretations are consistent with research that demonstrates that the quality of teacher-like feedback is often associated with the intentional framing of the pedagogical purpose of the feedback, including the clarity of instructional intent, the priority placed on the different components of the feedback, and the presence of constructive guidance rather than general comments [25], [45]. The framing of the role has also been identified as a critical factor in the design of prompts, with research suggesting that the role framing may influence the structure of the responses and the perceived utility of the responses, thus enabling the model behavior to align with educator-like communicative norms [40], [41].

Pedagogical mediation of teaching through prompt engineering appeared to be further supported by the sequencing of the feedback. The current study is consistent with recent research comparing the sequencing of feedback components in

AI-supported writing instruction, suggesting that the order of the components of the feedback may influence the type and potential uptake of the feedback [32], [63]; and that formative feedback tends to be more effective when framed from an improvement rather than a judgmental perspective [24], [66]. Although multiple prompt configurations yielded relatively high linguistic scores, the highest score was achieved by the combination of pedagogically grounded personas and task instructions that emphasized explanation and prioritization. Expert rationale suggests that effective language feedback is not only accurate but also accessible, clear in terms of explanation, and calibrated to the learner's proficiency, qualities that were more consistently present when the prompts were structured. This is consistent with recent empirical findings indicating that the quality of AI feedback improves when prompts are designed to promote personalization and pedagogical appropriateness, especially for non-native English speakers [2], [59], as well as broader meta-analytic evidence indicating that the relative benefits of AI feedback depend upon user framing and perception [13], [28].

Significant ethical considerations exist in the use of AI in culturally situated educational environments. A primary concern is the risk of inaccuracy; notably, the base prompts in this study were associated with a higher rate of hallucinations (approximately 55% higher than structured prompts), suggesting that poorly specified prompting conditions may lead to unreliable output. Other concerns include cultural bias and the possibility of over-reliance on AI, which may lead to a gradual erosion of teacher authority. To mitigate these risks, several safeguards were implemented: prompts were carefully constructed to restrict the AI persona (i.e., lecturer vs. cleric); inter-rater reliability checks were conducted (r = .89–.96); and AI outputs were systematically evaluated. These approaches are generally consistent with UNESCO's AI Ethics Framework [67] and emerging Islamic Technology Guidelines [60].

The incorporation of Islamic content into EFL writing feedback represents a potentially unique contribution of this study to the field. As English Medium Instruction in Religious Contexts attracts increasing scholarly attention, there is a growing need for frameworks for feedback that can both support academic argumentation and religiously-informed reasoning. The Islamic content ratings indicate that greater religious accuracy and cultural sensitivity tended to be associated with prompts that included pedagogical framing and contextual constraints, namely, lecturer personas and Islamic Higher Education framing. Expert and inter-rater commentary suggested that this was consistent with the notion that EFL instruction in religious contexts requires the negotiation of academic conventions and the preservation of Islamic epistemic commitments [49], [52], and is also consistent with Culturally Responsive Teaching frameworks suggesting that feedback must respect students' cultural and epistemological frames to be educationally legitimate [49], [68].

The data further suggest that instructional constraint may be an important factor in mitigating the risk of hallucination and output inconsistency. The combination of pedagogically grounded personas and feedback-only sequencing were associated with the highest levels of naturalness and reliability, suggesting that clearly-defined roles and task instructions that focus on explanation and prioritization may enhance the coherence and logical consistency of AI-generated output. These patterns are consistent with studies examining the reliability of LLMs in evaluative roles, which emphasize the importance of verification, consistency-checking, and careful framing [36], [37]. Further research on the limitations of LLMs also suggests that hallucinations and output instability are context-dependent and may be exacerbated by underspecified prompts, thus supporting the interpretation that prompt design presents a promising practical means of improving the reliability of educational applications of LLMs [1], [26].

The findings suggest that the most effective prompt strategy was not a single high-performance strategy in a particular domain, but rather a strategy that maximized balanced performance across pedagogy, language, Islamic content, and reliability. This pattern is consistent with recent frameworks conceptualizing effective AI feedback as multi-faceted, i.e., encompassing instructional value, personalization, and trustworthiness, rather than maximizing performance in a single dimension at the expense of others [12], [42]. Expert and inter-rater justification further suggested that balanced prompts may support coherent scaffolding while limiting epistemic risk, a particular concern in religiously-grounded writing contexts, where misinformed claims may have potential consequences for ethics and theology [27], [54].

In aggregate, these findings are consistent with the interpretation that prompt engineering serves as a form of pedagogical mediation, providing the LLM with instructional direction on how to perform feedback roles, determine priorities, and manage epistemic boundaries. The convergence of the quantitative rankings and expert/inter-rater justification strengthens the plausibility of this interpretation and suggests that prompt design should be viewed as a core instructional decision in the responsible use of generative AI, rather than as a secondary, technical adjustment. This conclusion is consistent with recent developments in the literature on AI Feedback Literacy, which argue that the successful and ethical application of generative AI will depend in part on the ability of educators to design, interpret, and regulate AI feedback practices in accordance with their pedagogical objectives [14], [18].

The implications of these findings for future research, teaching practice, and institutional policy are substantial. For research, the findings suggest that prompt engineering should be viewed as a basic element of educational design, not just a technical configuration. Recent large-scale and meta-analytic studies have demonstrated considerable variability in the effectiveness of AI-generated feedback based on context and implementation strategy [12], [13], [69], and the findings of this study further underscore the call for more detailed comparative analysis of design conditions, rather than simply contrasting AI with human feedback. The expert-validated framework developed in this study may provide a useful starting point for such analyses.

For teaching practice, the findings provide initial evidence that EFL instructors using generative AI in writing instruction may benefit from using Prompt 4 as the best configuration identified in this study. Prompts that define the AI as an

English lecturer and require feedback before grading appear to align with widely accepted principles of learner-centered and improvement-focused feedback [45], [66], and provide some support for recent calls for developing AI feedback literacy among educators, defined not simply as the ability to use tools, but as the ability to design AI-based prompts aligned with their pedagogical philosophy and assessment-for-learning objectives [14], [18]. EFL instructors in Islamic higher education may also benefit from implementing a hybrid model of using AI to generate initial draft feedback that the teacher then revises to ensure that the Islamic content is both correct and contextually appropriate, a safeguard that appears to be of particular importance given recent concerns regarding the limitations of generative AI in addressing the ethical and theological requirements of religious content [26], [56].

At the institutional level, the findings suggest that the responsible use of AI in education may require the creation of clear guidelines for the design, evaluation, and monitoring of AI-based prompts. Rather than either banning or unconditionally accepting generative AI, institutions may benefit from creating pedagogically-based prompt templates and expert-informed frameworks for evaluating the quality, reliability, and cultural relevance of AI-generated feedback [26], [65].

There are a number of limitations to this study that should be acknowledged. The sample size of four to five expert raters drawn from a specialized population, EFL instruction in an Islamic higher education context — limits the generalizability of the findings. Although the high inter-rater reliability (r = .89–.96) supports the internal validity of the findings, the perceptions of the students regarding the usefulness of the feedback in a classroom environment were not investigated. Future research may seek to build on these findings through quasi-experimental studies (Phase 2) using approximately 90 students, and applying the highest performing prompts, such as Prompt 4, in authentic classroom environments to investigate their effects on student writing and engagement.

VI. Conclusion

The findings of this research study clearly indicate that Prompt 4 (the English lecturer's reply to the student's assignment; feedback only), was the best overall scoring prompt among the tested prompting strategies, and therefore has great potential for developing outputs that are well-balanced and of high quality when evaluating the student assignments from different aspects. There were also some other relatively good-scoring prompts, like P6 and P7, which have shown to perform consistently. However, the results of the statistical analyses show that the variations of performance among the different prompting strategies are statistically insignificant (p=0.109). Therefore, the variations of performance should be considered descriptive, and not be taken as proof that one strategy performs better than another.

Quantitative results, supported by both expert and inter-rater rationalization indicated that two specific prompt engineering strategies, namely using persona-based prompts, and structuring the task sequence to provide students feedback first resulted in feedback that was evaluated to be more pedagogically effective, linguistically accurate, religiously appropriate and technologically reliable than either base prompts or those in which the LLM received feedback second.

Although the findings of this study have strong validation from the experts that participated in the study; further research will be required to confirm these findings in a classroom environment. Additionally, future research may be able to provide an empirical template for deploying AI-based feedback responsibly. In future research, a fully crossed repeated-measures experiment, where each prompt is used to write essays on the same set of topics will be applied. This will allow us to tightly control the effects of the task, and thereby make stronger claims about causality.

References

[1] P. Peykani, F. Ramezanlou, C. Tanasescu, and S. Ghanidel, “Large Language Models: A Structured Taxonomy and Review of Challenges, Limitations, Solutions, and Future Directions,” Applied Sciences (Switzerland), vol. 15, no. 14, p. 8103, 2025, doi: 10.3390/app15148103.

[2] M. Jovic, S. Papakonstantinidis, and R. Kirkpatrick, “From red ink to algorithms: investigating the use of large language models in academic writing feedback,” Language Testing in Asia, vol. 15, no. 1, p. 59, 2025, doi: 10.1186/s40468-025-00389-2.

[3] J. M. Gayed, M. K. J. Carlon, A. M. Oriola, and J. S. Cross, “Exploring an AI-based writing Assistant’s impact on English language learners,” Computers and Education: Artificial Intelligence, vol. 3, p. 100055, 2022, doi: 10.1016/j.caeai.2022.100055.

[4] M. Ekizoğlu and A. N. Demir, “The role of AI assisted writing feedback in developing secondary students writing skills,” Discover Education, vol. 4, no. 1, p. 454, 2025, doi: 10.1007/s44217-025-00919-3.

[5] Marzuki, U. Widiati, D. Rusdin, Darwin, and I. Indrawati, “The impact of AI writing tools on the content and organization of students’ writing: EFL teachers’ perspective,” Cogent Education, vol. 10, no. 2, 2023, doi: 10.1080/2331186X.2023.2236469.

[6] A. Alnemrat, H. Aldamen, M. Almashour, M. Al-Deaibes, and R. AlSharefeen, “AI vs. teacher feedback on EFL argumentative writing: a quantitative study,” Front. Educ. (Lausanne)., vol. 10, 2025, doi: 10.3389/feduc.2025.1614673.

[7] K. Hyland and F. Hyland, “Feedback on second language students’ writing,” Language Teaching, vol. 39, no. 2, pp. 83–101, 2006, doi: 10.1017/S0261444806003399.

[8] N. Eryilmaz, A. I. Kennedy, R. Strietholt, and S. Johansson, “Teacher job satisfaction: International evidence on the role of school working conditions and teacher characteristics,” Studies in Educational Evaluation, vol. 86, p. 101474, 2025, doi: 10.1016/j.stueduc.2025.101474.

[9] H. Wang, Y. Sun, W. Wang, and H. Liang, “Exploring the relationship between teachers’ perceived workload, challenge-hindrance stress, and work engagement: a person-centered approach,” BMC Psychol., vol. 13, no. 1, 2025, doi: 10.1186/s40359-025-02537-y.

[10] M. Asadi, S. Ebadi, and L. Mohammadi, “The impact of integrating ChatGPT with teachers’ feedback on EFL writing skills,” Think. Skills Creat., vol. 56, p. 101766, 2025, doi: 10.1016/j.tsc.2025.101766.

[11] C. Yang, J. Shao, and Y. Guo, “Can teacher feedback be substituted by Gen-AI? A comparative study of writing scores, feedback characteristics, revision and motivation,” in Innovation in Language Learning and Teaching, 2025, pp. 1–24. doi: 10.1080/17501229.2025.2609877.

[12] S. Ba, L. Yang, Z. Yan, C. K. Looi, and D. Gašević, “Unraveling the mechanisms and effectiveness of AI-assisted feedback in education: A systematic literature review,” Computers and Education Open, vol. 9, p. 100284, 2025, doi: 10.1016/j.caeo.2025.100284.

[13] R. Kaliisa, K. Misiejuk, S. López-Pernas, and M. Saqr, How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions. Educational Psychology, 2025. doi: 10.1080/01443410.2025.2553639.

[14] K. Liu and F. D. Deris, “AI feedback literacy in higher education: understanding, measuring, and predicting student feedback uptake,” in Assessment and Evaluation in Higher Education, 2025, pp. 1–15. doi: 10.1080/02602938.2025.2587924.

[15] T. Nazaretsky, H. Gabbay, and T. Käser, “Can students judge like experts? A large-scale study on the pedagogical quality of AI and human personalized formative feedback,” Computers and Education: Artificial Intelligence, vol. 10, p. 100533, 2026, doi: 10.1016/j.caeai.2025.100533.

[16] L. C. Chen, H. T. Weng, M. S. Pardeshi, C. M. Chen, R. K. Sheu, and K. C. Pai, “Evaluation of Prompt Engineering on the Performance of a Large Language Model in Document Information Extraction,” Electronics (Switzerland), vol. 14, no. 11, p. 2145, 2025, doi: 10.3390/electronics14112145.

[17] S. Lee and S. Kang, “The influence of generative AI with prompt engineering on creative design in architectural education,” Journal of Asian Architecture and Building Engineering, vol. 1–16, 2025, doi: 10.1080/13467581.2025.2552446.

[18] N. Knoth, A. Tolzin, A. Janson, and J. M. Leimeister, “AI literacy and its implications for prompt engineering strategies,” Computers and Education: Artificial Intelligence, vol. 6, 2024, doi: 10.1016/j.caeai.2024.100225.

[19] 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 (Switzerland), vol. 6, no. 2, p. 35, 2025, doi: 10.3390/ai6020035.

[20] E. Ryen, “Klafki’s critical-constructive Didaktik and the epistemology of critical thinking,” Journal of Curriculum Studies, vol. 52, no. 2, pp. 214–229, 2020, doi: 10.1080/00220272.2019.1657959.

[21] H. Seyri and F. Ghiasvand, “‘Teaching is basically feeling’: Unpacking EFL Teachers’ perceived emotions and regulatory strategies in AI-Powered L2 speaking and writing skills instruction,” Computers and Education Open, vol. 8, p. 100264, 2025, doi: 10.1016/j.caeo.2025.100264.

[22] M. Mekheimer, “Generative AI-assisted feedback and EFL writing: a study on proficiency, revision frequency and writing quality,” Discover Education, vol. 4, no. 1, p. 170, 2025, doi: 10.1007/s44217-025-00602-7.

[23] Z. Zhang, S. Aubrey, X. Huang, and T. K. F. Chiu, “The role of generative AI and hybrid feedback in improving L2 writing skills: a comparative study,” in Innovation in Language Learning and Teaching, 2025, pp. 1–19. doi: 10.1080/17501229.2025.2503890.

[24] R. Yu and L. Yang, “ESL/EFL Learners’ Responses to Teacher Written Feedback: Reviewing a Recent Decade of Empirical Studies,” Front. Psychol., vol. 12, 2021, doi: 10.3389/fpsyg.2021.735101.

[25] H. S. Sanchez and L. de A. D. Rodrigues, “Pedagogical intentions behind teacher written feedback: The perspectives and practices of an English language teacher educator in Argentina,” J. Engl. Acad. Purp., vol. 69, p. 101370, 2024, doi: 10.1016/j.jeap.2024.101370.

[26] I. M. García-López and L. Trujillo-Liñán, “Ethical and regulatory challenges of Generative AI in education: a systematic review,” Front. Educ. (Lausanne)., vol. 10, 2025, doi: 10.3389/feduc.2025.1565938.

[27] C. Papakostas, “Artificial Intelligence in Religious Education: Ethical, Pedagogical, and Theological Perspectives,” Religions (Basel)., vol. 16, no. 5, p. 563, 2025, doi: 10.3390/rel16050563.

[28] W. Wang et al., “The Effectiveness of AI-Supported Personalized Feedback on Students’ Learning Outcomes and Motivation: A Meta-Analysis,” Journal of Educational Computing Research, vol. 0, no. 0, 2025, doi: 10.1177/07356331251410020.

[29] A. M. Vieriu and G. Petrea, “The Impact of Artificial Intelligence (AI) on Students’ Academic Development,” Educ. Sci. (Basel)., vol. 15, no. 3, p. 343, 2025, doi: 10.3390/educsci15030343.

[30] J. Meyer et al., “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, vol. 6, 2024, doi: 10.1016/j.caeai.2023.100199.

[31] R. Khan, M. T. Qamar, M. S. Ansari, and J. Yasmeen, “Enhancing or impairing? Exploring Indian EFL learners’ academic writing narratives with ChatGPT,” Cogent Education, vol. 12, no. 1, 2025, doi: 10.1080/2331186X.2025.2514329.

[32] T. T. T. Tran, “Enhancing EFL Writing Revision Practices: The Impact of AI- and Teacher-Generated Feedback and Their Sequences,” Educ. Sci. (Basel)., vol. 15, no. 2, p. 232, 2025, doi: 10.3390/educsci15020232.

[33] T. Heydarnejad, “Unmasking the impacts of self-evaluation in AI-supported writing instruction on EFL learners’ emotion regulation, self-competence, motivation, and writing achievement,” Computers and Education: Artificial Intelligence, vol. 9, p. 100494, 2025, doi: 10.1016/j.caeai.2025.100494.

[34] B. Otaki, S. Naganuma, T. Jin, and J. Oshima, Differences in perceptions of generative AI feedback by non-native English-speaking students and educators in higher education. Educational Psychology, 2025. doi: 10.1080/01443410.2025.2564893.

[35] R. F. Kizilcec et al., “Perceived impact of generative AI on assessments: Comparing educator and student perspectives in Australia, Cyprus, and the United States,” Computers and Education: Artificial Intelligence, vol. 7, p. 100269, 2024, doi: 10.1016/j.caeai.2024.100269.

[36] Y. Sawaki, Y. Ishii, H. Yamada, and T. Tokunaga, “Examining the consistency of instructor versus large language model ratings on summary content: Toward checklist-based feedback provision with second language writers,” Language Testing, vol. 42, no. 4, pp. 447–475, 2025, doi: 10.1177/02655322251349217.

[37] H. Seo et al., “Large Language Models as Evaluators in Education: Verification of Feedback Consistency and Accuracy,” Applied Sciences (Switzerland), vol. 15, no. 2, p. 671, 2025, doi: 10.3390/app15020671.

[38] J. Lo, C. Wong, A. Ng, P. Wong, D. Cheung, and P. Lai, “Stretching AI’s reach: Assessing an AI-driven feedback system for extended academic writing,” Computers and Education: Artificial Intelligence, vol. 10, p. 100511, 2026, doi: 10.1016/j.caeai.2025.100511.

[39] M. Gozzi and F. Di Maio, “Comparative Analysis of Prompt Strategies for Large Language Models: Single-Task vs. Multitask Prompts,” Electronics (Switzerland), vol. 13, no. 23, p. 4712, 2024, doi: 10.3390/electronics13234712.

[40] A. P. Correia, S. Hickey, and F. Xu, “Realizing the possibilities of the large language models: Strategies for prompt engineering in educational inquiries,” Theory Pract., vol. 64, no. 4, pp. 434–447, 2025, doi: 10.1080/00405841.2025.2528545.

[41] Y. Choi, M. Lee, S. Han, and J. Han, “Effects of Prompt Elements on Problem-Solving Performance and User Experience: Insights from ChatGPT Interactions,” Sage Open, vol. 15, no. 4, 2025, doi: 10.1177/21582440251381680.

[42] Y. Aperstein, Y. Cohen, and A. Apartsin, “Generative AI-Based Platform for Deliberate Teaching Practice: A Review and a Suggested Framework,” Educ. Sci. (Basel)., vol. 15, no. 4, p. 405, 2025, doi: 10.3390/educsci15040405.

[43] D. Lu and Y. Zeng, “Exploring the use of ChatGPT-generated model texts as a feedback instrument: EFL students’ text quality and perceptions,” in Innovation in Language Learning and Teaching, 2025, pp. 1–21. doi: 10.1080/17501229.2025.2525341.

[44] N. Lo, A. Wong, and S. Chan, “The impact of generative AI on essay revisions and student engagement,” Computers and Education Open, vol. 9, p. 100249, 2025, doi: 10.1016/j.caeo.2025.100249.

[45] A. A. Aldino et al., “Analytics of Learner-Centered Feedback: A Large-Scale Case Study in Higher Education,” Comput. Educ., vol. 237, p. 105360, 2025, doi: 10.1016/j.compedu.2025.105360.

[46] M. McGuire, “Feedback, reflection and psychological safety: rethinking assessment for student well-being in higher education,” in Assessment and Evaluation in Higher Education, 2025, pp. 1–25. doi: 10.1080/02602938.2025.2548590.

[47] B. Huang, S. Xie, and T. K. F. Hew, “Effects of integrating generative AI, teacher, and peer feedback on Middle school students’ attitudes, writing proficiency, and peer-assessment scoring ability: a quasi-experimental study,” Educ. Psychol. (Lond)., vol. 1–23, 2026, doi: 10.1080/01443410.2025.2611980.

[48] S. Monica, S. Arsyad, and B. Waluyo, “English medium instruction in Islamic higher education: Challenges of Lecturer readiness in Indonesia,” Social Sciences and Humanities Open, vol. 12, p. 101900, 2025, doi: 10.1016/j.ssaho.2025.101900.

[49] H. P. Ilyas and W. Tarmini, “Interpreting pedagogical beliefs: English language teaching in Islamic educational institutions,” Englisia: Journal of Language, Education, and Humanities, vol. 13, no. 1, pp. 207–225, 2025, doi: 10.22373/ej.v13i1.32156.

[50] N. Asnawi and M. Zuhdi, “Deconstructing Logocentrism and School-Centrism in Indonesia’s Islamic Education: A Critical Epistemological Analysis,” Educ. Sci. (Basel)., vol. 15, no. 12, p. 1615, 2025, doi: 10.3390/educsci15121615.

[51] S. Syamsuri and A. Musgamy, “Strengthening Student Character Through Qur’anic Values in Islamic Education,” Journal La Edusci, vol. 6, no. 6, pp. 1246–1254, 2026, doi: 10.37899/journallaedusci.v6i6.2845.

[52] A. B. Selim, S. Sumaya, and Z. F. Chowdhury, “Teaching English While Preserving Islamic Tradition: A Balanced Educational Approach,” Journal of Applied and Action Research in Islamic Education, vol. 1, no. 1, pp. 58–71, 2025, doi: 10.70771/jaarie.v1i1a4.

[53] S. Supriadi, H. Hosaini, and Z. H. Sain, “Transformation of Literacy-Based Islamic Education Learning Management Integration in Elementary Schools,” International Journal of Social Learning (IJSL), vol. 5, no. 1, pp. 294–304, 2024, doi: 10.47134/ijsl.v5i1.358.

[54] R. Tsuria and Y. Tsuria, “Artificial Intelligence’s Understanding of Religion: Investigating the Moralistic Approaches Presented by Generative Artificial Intelligence Tools,” Religions (Basel)., vol. 15, no. 3, p. 375, 2024, doi: 10.3390/rel15030375.

[55] J. C. Tom, T. W. Ferguson, and B. C. Martinez, “Religion and Racial Bias in Artificial Intelligence Large Language Models,” 2025. doi: 10.1177/23780231251377210.

[56] M. Andok, Z. Rajki, and S. Dornics, “The Use of Artificial Intelligence Tools for Religious Purposes: Empirical Research Among Hungarian Religious Communities,” Religions (Basel)., vol. 16, no. 8, p. 999, 2025, doi: 10.3390/rel16080999.

[57] F. Khodabandeh, “The Impact of AI-Assisted vs. Instructor-Provided Feedback on Awe, Autonomy, and Letter Writing Skills in Introverted and Extroverted EFL Learners in Online Learning Environments,” Journal of Educational Computing Research, vol. 0, no. 0, 2025, doi: 10.1177/07356331251412228.

[58] A. S. Nelson, P. V. Santamaría, J. S. Javens, and M. Ricaurte, “Students’ Perceptions of Generative Artificial Intelligence (GenAI) Use in Academic Writing in English as a Foreign Language †,” Educ. Sci. (Basel)., vol. 15, no. 5, p. 611, 2025, doi: 10.3390/educsci15050611.

[59] F. Algobaei and E. Alzain, “Prompt engineering for non-native English learners: A generative AI approach to personalised language feedback,” Social Sciences and Humanities Open, vol. 13, p. 102341, 2026, doi: 10.1016/j.ssaho.2025.102341.

[60] M. J. Shamsuddin, M. F. Sawari, and F. Karim, “Islamic guidelines for content creators: A jurisprudential framework,” Ulum Islamiyyah, vol. 37, no. 03, pp. 46–63, 2025, doi: 10.33102/uij.vol37no03.720.

[61] M. Stahl, L. Biermann, A. Nehring, and H. Wachsmuth, “Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation,” in Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, 2024, pp. 283–298. [Online]. Available: http://arxiv.org/abs/2404.15845

[62] K. Hyland and F. Hyland, “Contexts and issues in feedback on L2 writing,” in Feedback in Second Language Writing: Contexts and Issues, K. Hyland and F. Hyland, Eds., Cambridge University Press, 2019, pp. 1–22. doi: 10.1017/9781108635547.003.

[63] J. I. Hoyos Valdiviezo, J. C. Dos Santos, and R. M. Vinueza Beltrán, “Effectiveness of Accurate ChatGPT Commands in Enhancing A2 Level Academic Writing Feedback,” Forum for Linguistic Studies, vol. 7, no. 12, pp. 12–25, 2025, doi: 10.30564/fls.v7i12.11647.

[64] S. Ba, L. Yang, Z. Yan, C. K. Looi, and D. Gašević, “Unraveling the mechanisms and effectiveness of AI-assisted feedback in education: A systematic literature review,” Computers and Education Open, vol. 9, p. 100284, 2025, doi: 10.1016/j.caeo.2025.100284.

[65] J. Venter, S. A. Coetzee, and A. Schmulian, “Exploring the use of artificial intelligence (AI) in the delivery of effective feedback,” Assess. Eval. High. Educ., vol. 50, no. 4, pp. 516–536, 2025, doi: 10.1080/02602938.2024.2415649.

[66] G. T. L. Brown, C. Andersson, M. Winberg, B. Palmberg, and T. Palm, “Teacher conceptions of assessment and feedback predicting formative feedback practices: helping students to not ignore improvement-oriented feedback,” Scandinavian Journal of Educational Research, pp. 1–19, 2025, doi: 10.1080/00313831.2025.2468183.

[67] G. Ramos, “UNESCO’s Recommendation on the Ethics of Artificial Intelligence (2021),” International Organization Initiatives, pp. 203–214, 2025, doi: 10.1093/9780197803325.003.0020.

[68] J. A. Ogodo, “Culturally Responsive Pedagogical Knowledge: An Integrative Teacher Knowledge Base for Diversified STEM Classrooms,” Educ. Sci. (Basel)., vol. 14, no. 2, p. 124, 2024, doi: 10.3390/educsci14020124.

[69] P. Wei, X. Wang, and H. Dong, “The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: a randomized controlled trial,” Front. Psychol., vol. 14, 2023, doi: 10.3389/fpsyg.2023.1249991.