

RESEARCH ARTICLE

Designing an On-Premise, RAG-Based Rater Training Program for Korean Language Arts Essay Assessment

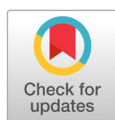
Jinhwan Son*

Korea National University of Education

*Corresponding Author: Jinhwan Son, cdere0903@naver.com

ABSTRACT

The expansion of descriptive and essay-type assessment in South Korea's 2022 Revised National Curriculum and 2028 College Admissions Reform Plan has intensified the need for reliable scoring by Korean Language Arts teachers. Existing rater training approaches face a persistent dual limitation: face-to-face workshops are constrained by spatiotemporal barriers and their one-off nature, while computer-based systems provide only superficial feedback that fails to address the cognitive processes underlying scoring decisions. This study proposes Retrieval-Augmented Generation (RAG) as a technology capable of bridging this gap and presents the design of a RAG-based rater training program. Drawing on Cognitive Apprenticeship, Frame-of-Reference training, and Deliberate Practice theory, six core design principles were derived through a two-axis analysis integrating theoretical mechanisms with design limitations identified in prior programs. These principles were translated into a seven-stage program following a 2-Round learning path (pretest–intervention–posttest). The program's RAG system, built on open-source Korean-language LLMs (EXAONE 3.5 7.8B and SOLAR 10.7B) deployed on institutional on-premise servers, differentiates the roles of retrieval and generation across stages: RAG ensures knowledge consistency through precise case retrieval from an expert scoring database, while the LLM performs complex reasoning and personalized feedback synthesis. An adaptive feedback mechanism adjusts content and tone based on diagnosed scoring tendencies (leniency or severity), and a structured What-Where-Why scoring rationale requirement in the second scoring round promotes metacognitive reflection by making the cognitive process of scoring—not merely its numerical outcome—a target of training and feedback. As a design-stage study, this work presents the program's architecture in implementable detail; empirical effectiveness verification through field deployment with practicing teachers is reserved for subsequent research.



OPEN ACCESS

Brain, Digital, & Learning
2026, Vol. 16, No. 3, 341–359

<https://doi.org/10.31216/BDL.2026.16.3.8>

Received: July 14, 2026

Revised: September 29, 2026

Accepted: September 29, 2026

© 2026 Institute of Brain·AI based Education,
Korea National University of Education



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Keywords: rater training, scoring reliability, retrieval-augmented generation, Korean language arts, essay-type assessment, cognitive apprenticeship, large language model

Introduction

The expansion of descriptive and essay-type assessment has become a central pillar of educational reform in South Korea. The 2028 College Admissions Reform Plan (Ministry of Education, 2023) explicitly calls for an increased share of essay-type items that measure higher-order thinking, and the 2022 Revised National Curriculum (Ministry of Education, 2022) emphasizes process-centered assessment in which students construct, express, and justify their own ideas. This aligns with the international shift from selection-type testing toward the assessment of complex cognitive competencies (OECD, 2018).

However, the benefits of essay-type assessment depend on scoring reliability. Because constructed responses have no single correct answer, every score relies on the rater's professional judgment, and when this judgment lacks intra-rater and inter-rater consistency, the validity of the assessment is undermined. This problem is particularly acute in Korean Language Arts. For the same task, students produce diverse arguments and organizational structures, and raters apply rubrics differently according to their interpretive frames (Park, J., 2014); moreover, the scoring criteria require the simultaneous evaluation of content, organization, and expression, while the competency-based rubrics of the revised curriculum resist simple right-or-wrong judgments. Furthermore, the systematic rater errors that Myford and Wolfe (2004) identified through many-facet Rasch measurement—leniency/severity, central tendency, and halo effects—are not random mistakes but stable cognitive tendencies that persist unless explicitly addressed through training.

Rater training is the most fundamental mechanism for securing scoring reliability, yet existing approaches face a dual limitation. Traditional face-to-face workshops are effective at forming shared understanding through group discussion and expert-led calibration, but they are constrained by scheduling difficulties, geographic barriers, and their one-off nature (Jang, 2010; Lee, Y., 2015). Conversely, existing computer-based training systems overcome spatiotemporal constraints but remain limited to superficial feedback at the level of score comparison or right/wrong confirmation, failing to address the deep cognitive processes underlying scoring decisions (Choi, Y. et al., 2019; Elder et al., 2007). Neither approach alone can simultaneously remove accessibility barriers and provide context-sensitive, individualized feedback.

The design principles of the present program rest on three learning theories. Cognitive Apprenticeship (Collins et al., 1989) makes an expert's tacit thinking visible to learners through the instructional methods of modeling, coaching, scaffolding, articulation, reflection, and exploration. Expert scoring entails a cognitive process that is invisible to novice raters—reading a response, activating the rubric, and comparing evidence to reach a judgment—and making this process explicit through expert rationales and self-explanation is a prerequisite for internalizing the criteria. Frame-of-Reference training (FOR; Bernardin & Buckley, 1981) exposes raters to benchmark responses at each score level to align their individual standards with expert consensus, and is especially well suited to domains in which multiple valid interpretations exist. Indeed, Weigle (1998) empirically demonstrated in writing assessment that benchmark calibration training improves raters' internal consistency. Deliberate Practice (Ericsson, 2006) holds that expert performance develops through well-defined tasks, immediate feedback, and repetition with correction, suggesting the need for repeated scoring practice accompanied by diagnostic feedback rather than a single workshop. In sum, Cognitive Apprenticeship provides the instructional sequence, Frame-of-Reference training specifies the expert consensus toward

which learners must converge, and Deliberate Practice defines the conditions under which convergence occurs—target tasks, immediate feedback, and repetition. The present program integrates all three theories.

As the technology for implementing these three principles in a scalable digital environment, this study proposes Retrieval-Augmented Generation (RAG). RAG is an architecture that, for a given user query, retrieves relevant documents from a curated knowledge database and provides them as context to a large language model (LLM) to generate a response (Lewis et al., 2020), reducing the risk of LLM hallucination by grounding generation in the retrieved evidence. Applied to rater training, RAG offers three benefits. First, because it draws on a database containing expert scoring cases, rubric annotations, and anchor-paper rationales, it maintains responses aligned with expert consensus regardless of how a learner formulates a query. Second, by matching a learner's scoring patterns through semantic similarity, it presents the expert cases most relevant to each learner's weaknesses. Third, by presenting the retrieved evidence together with feedback, it makes the source of every recommendation traceable to specific expert cases. RAG thus implements the three aforementioned theories in a single environment by modeling expert thinking, aligning learners' judgments with expert anchors, and providing individualized immediate feedback. Because AI-based assessment tools are more effective as a complement to human judgment than as a replacement for it (Ramesh & Sanampudi, 2022), however, the present program positions RAG not as an automated scorer but as an intelligent training environment that strengthens raters' expertise.

The purpose of this study is to design a RAG-based rater training program for Korean Language Arts essay assessment. Specifically, it addresses the following two research questions.

1. What are the design principles and structural composition of a RAG-based rater training program?
2. What are the program's stage-by-stage operating principles and its strategy for using RAG and the LLM?

This study corresponds to the design stage of a larger developmental research effort. It derives design principles from theoretical analysis and a review of prior cases and presents the program's architecture, learning path, and feedback mechanism at an implementable level of detail; actual system development, field application, and empirical effectiveness verification are left as subsequent tasks.

Methods

Research Procedure and Materials

This study followed a four-stage research procedure. In Stage 1, a needs analysis and literature review were conducted to identify the components of rater expertise, examine existing rater training programs, and explore the pedagogical applicability of RAG systems. In Stage 2, six core design principles were derived by integrating insights from Cognitive Apprenticeship theory (Collins et al., 1989), Frame-of-Reference training (Bernardin & Buckley, 1981), and Deliberate Practice theory (Ericsson, 2006) with the design limitations identified in prior rater training research (Choi, Y. et al., 2019; Park, J., 2013; Jang, 2010). In Stage 3, these principles were translated into the program's overall structure, comprising a 2-Round learning path, module-level learning objectives, and the RAG system architecture. In Stage 4, the operating procedures for each training stage, the RAG-based feedback mechanism, and the program evaluation scheme were

specified in detail.

The assessment materials used in this study were drawn from the AI-Hub “Essay Writing Assessment Data” constructed and released by the National Information Society Agency (2021). The dataset comprises approximately 50,000 student essays from elementary, middle, and high school students on 50 topics designed with expert consultation, including argumentative writing such as making claims, arguing for or against a position, and proposing alternatives, as well as expository essays. Each essay was rated by three assessment experts using an analytic rubric of 11 elements spanning the three dimensions of expression, organization, and content, and the labels record each rater's scores and the average score in a structured format.

This dataset was selected for three reasons. First, because it consists of authentic student writing in Korean, it secures the ecological validity of Korean Language Arts rater training. Second, because each case includes ratings from three experts, the reliability of the scoring results can be verified, and inter-rater agreement can be used as a selection criterion for inclusion in the knowledge database. Third, the element-wise analytic rubric structure directly aligns with the program's element-level scoring and feedback design. This dataset is, however, a general writing-assessment resource that does not cover the full range of Korean Language Arts essay assessment, such as descriptive responses in the reading and literature domains; the program is therefore designed around argumentative-writing assessment, with extension to other domains left as a future task.

This study also examined another assessment resource provided through the same source, AI-Hub—the “Essay-Type Writing Assessment Data” (Table 1). The latter is a more recent dataset constructed in 2024 that explicitly designates itself as “essay-type” and offers prompts grounded in the Korean, social studies, and science subjects. Nevertheless, this study selected the “Essay Writing Assessment Data” because of two methodological requirements imposed by the program's core design. First, the program's knowledge database uses the agreement of three raters per case as the criterion for inclusion, which requires independent ratings from multiple experts for each case; the “Essay Writing Assessment Data” provides three-rater ratings, whereas the “Essay-Type Writing Assessment Data” uses a two-rater scheme. Second, the program's element-level scoring and feedback and its leniency-versus-severity demonstration presuppose a fine-grained distribution of element-wise scores across expression, organization, and content; the “Essay Writing Assessment Data” provides analytic scores for up to 11 elements per rater, directly meeting this design requirement. In addition, the “Essay Writing Assessment Data” consists of authentic student writing across all grade bands from elementary to high school, whereas the “Essay-Type Writing Assessment Data” does not include high school materials and contains some outputs produced by generative models. The “Essay Writing Assessment Data” was constructed in 2021, however, making it a relatively older resource that encompasses essays in general rather than being limited to “essay-type” writing; extension to more recent resources such as the “Essay-Type Writing Assessment Data” is therefore left as a future task.

Table 1. Comparison of candidate AI-Hub datasets

Criterion	Essay Writing Assessment Data (selected)	Argumentative Writing Assessment Data
Build year	2021	2024
Labeled volume	~50,000 essays	20,000 essays
Grade range	Elementary 4–High 3 (all bands)	Elementary 5–6, Middle 1–3 (no high school)
Raters per essay	3 experts	2 (scoring)
Rubric structure	Up to 11 analytic elements (expression / organization / content)	8 topics (task / content / organization / expression)
Per-rater element scores	Provided (3 raters + average)	Analytic scores
Data origin	Actual student writing	Partly LLM-generated

Expert Knowledge Database Construction

The RAG system's knowledge database is constructed through a two-stage procedure that attaches experts' narrative scoring rationales to cases selected from the above dataset. In the first stage, case selection, only cases in which the three raters' element-wise scores fully agreed or converged within an adjacent score (± 1) are selected as candidates for inclusion, based on the deviation among the three raters' element-wise scores. This ensures that the scoring results recorded in the knowledge database reflect expert consensus rather than the idiosyncratic judgment of an individual rater.

In the second stage, rationale attachment, two in-service Korean Language Arts teachers with experience in developing and scoring essay-type assessment items are commissioned as expert raters to perform, for the selected cases, the writing of element-wise scoring rationales in the What-Where-Why structure, the selection of anchor papers by score level, the designation of borderline cases, and the description of rubric annotations and item intent. The two raters compose each output independently and then cross-review one another's work, with any items of disagreement resolved into a consensus version through deliberation. In this procedure, the dataset's original ratings are not altered, and the role of the expert raters is limited to making explicit the expert reasoning behind the established scores. Score reliability is thus secured by the dataset's three-rater ratings and agreement-based selection, while the expertise of the rationales is secured by the two raters' qualification criteria and consensus procedure.

Language Models and Deployment Environment

Two open-source Korean-language LLMs were selected for the RAG system. The primary model is EXAONE 3.5 7.8B (LG AI Research, 2024), and the fallback model is SOLAR 10.7B (Kim, D. et al., 2024). The most important selection criterion was data privacy. Essay-type responses contain students' thought processes, personal values, and personal experiences, and teachers' scoring data, being directly tied to assessment outcomes, constitute sensitive information; all of these fall under the regulation of the Personal Information Protection Act. Transmitting such data through external API services such as OpenAI or Anthropic would route it to overseas servers, conflicting with the institutional data-sovereignty requirements of Korean educational settings. Both selected models were developed by Korean companies and released under open-source licenses, allowing deployment on on-premise servers managed by a school or an office of education. The present design presupposes such an institution-managed on-premise environment, an assumption made to satisfy data-sovereignty requirements. This architecture ensures that all data remain within the institution while allowing teachers to connect through a web browser without spatiotemporal constraints.

To maximize operational efficiency on institutional servers, GGUF 4-bit quantization (Q4_K_M) was applied. Under this configuration, EXAONE 3.5 7.8B requires approximately 5–6 GB and SOLAR 10.7B approximately 7–8 GB of VRAM, enabling stable operation without expensive dedicated GPU infrastructure. In small-scale settings such as an individual school, local installation on a teacher's personal computer is also a feasible alternative.

These mid-sized models score lower than frontier API models on general-purpose benchmarks, and the present design rests on the assumption that the RAG architecture can substantially compensate for this gap. Because the system provides accurate context retrieved from a curated knowledge database—expert scoring cases, rubric annotations, and anchor-paper rationales—the LLM's primary task is not to generate answers from its own parametric knowledge but to faithfully synthesize and reorganize the provided context. Although retrieval augmentation has been shown to improve

the factual consistency of generation on knowledge-intensive tasks (Lewis et al., 2020), whether mid-sized Korean models achieve sufficient quality on this program's feedback-generation task remains to be confirmed through technical demonstration and subsequent evaluation. Both models exhibit strong Korean-language performance, and EXAONE 3.5 ranks among the top models on the KoMT-Bench and LogicKor benchmarks (LG AI Research, 2024).

The program is designed as a web-based online application operating on an institutional on-premise server. Teachers access the system through a standard web browser without installing additional software, consistent with the program's aim of overcoming the spatiotemporal constraints of conventional face-to-face rater training.

Results and Discussion

Design Principles

The six core design principles were derived through a two-axis analysis. In the first axis, core pedagogical mechanisms were extracted from the three theories reviewed above: the externalization of expert thinking from Cognitive Apprenticeship (Collins et al., 1989), consensus-based criterion alignment from Frame-of-Reference training (Bernardin & Buckley, 1981), and goal-directed practice accompanied by immediate feedback from Deliberate Practice (Ericsson, 2006), with prior discussions of assessment literacy (Park, Y., 2011; Stiggins, 1991) referenced as supporting grounds. In the second axis, the limitations of prior rater training programs were identified as requirements to be overcome: the spatiotemporal constraints of face-to-face workshops (Jang, 2010), the non-; implementation of systems that remained at the proposal stage (Park, J., 2013), and the superficial right/wrong-confirmation feedback of existing computer-based programs (Choi, Y. et al., 2019). The six principles derived from the cross-analysis of these two axes operate collaboratively toward effective rater training and are summarized in Table 2.

Table 2. Six core design principles

Principle	Description	Theoretical basis	Program implementation
1. Deepening assessment understanding	Raters must develop a thorough understanding of assessment items, scoring criteria, and evaluation objectives	Park, Y. (2011); Stiggins (1991)	RAG-based retrieval of item intent, rubric explanations, and expert rationales in response to rater queries
2. Progressive mastery	Training should progress from simple to complex tasks and from guided to independent performance	Collins et al. (1989); Ericsson (2006)	Adaptive learning path with scaffolding adjusted to learner progress (fading support as proficiency increases)
3. Immediate personalized feedback	Learners should receive timely, individualized feedback on their scoring performance	Ericsson (2006); Park, J. (2013)	Three-tier feedback (instant summary, detailed element-by-element analysis, scoring tendency diagnosis) generated via RAG+LLM
4. Criterion consistency	Raters must develop both intra-rater consistency and inter-rater consistency (aligning with expert consensus)	Bernardin & Buckley (1981); Jang (2010)	Calibration training with anchor papers, consistency metric visualization, and discrepancy detection
5. Expert knowledge systematization	Tacit expert knowledge must be made explicit and systematically accessible	Collins et al. (1989)	Expert scoring database with detailed rationales; semantic similarity search retrieves relevant expert cases
6. Self-regulated learning	Learners should be able to monitor their own scoring performance and direct their own improvement	Ericsson (2006)	Learning dashboard, self-assessment tools, and RAG-suggested learning paths

These six principles are interrelated rather than independent. Deepening assessment understanding is the foundation of all the others: unless raters clearly understand what a rubric measures, they cannot apply it consistently or respond meaningfully to feedback. Progressive mastery and immediate personalized feedback operate complementarily, with the system adjusting task difficulty according to a learner's demonstrated level while providing tailored feedback that addresses specific scoring gaps. Criterion consistency is realized through the systematization of expert knowledge, with the anchor papers and expert rationales in the knowledge database serving as concrete reference points for measuring and calibrating rater consistency. Finally, self-regulated learning is the overarching meta-principle, enabling learners to monitor their own progress across understanding, mastery, and consistency and to take ownership of their professional development.

The RAG system functions as an integrating technology that runs through all six principles. It retrieves domain-specific knowledge on demand, adjusts the level and type of support to a learner's current state, and provides data-driven consistency analysis and learning analytics for self-regulation, thereby supporting all six principles. This tight coupling of pedagogical design principles with the RAG architecture distinguishes the present program from prior approaches that treated training content and delivery technology as separate concerns.

System Architecture

The program adopts a four-layer modular architecture that separates pedagogical logic from technical infrastructure to facilitate independent maintenance and scalability. Fig. 1 depicts the overall architecture and the primary-fallback model strategy.

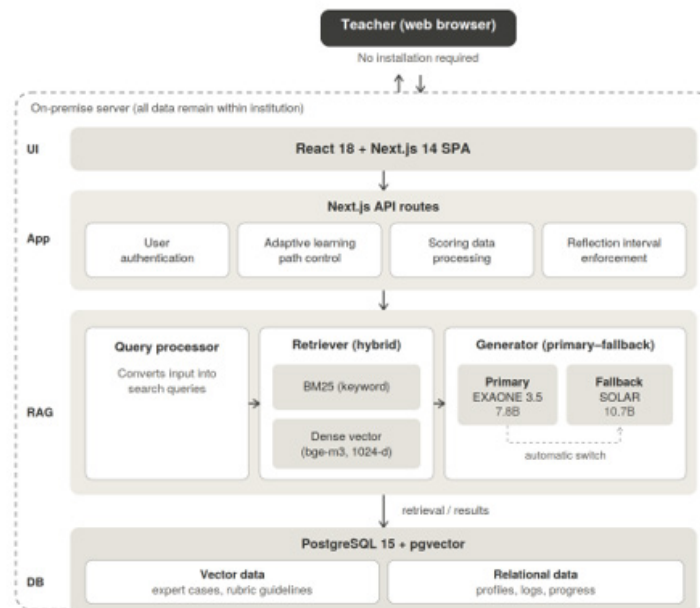


Fig. 1. System architecture and primary-fallback strategy

The user interface layer is implemented as a single-page application (SPA) based on React 18 and Next.js 14. The SPA design eliminates unnecessary page reloads when switching between scoring tasks and feedback screens, helping learners maintain cognitive flow throughout a training session.

The application layer, built on Next.js API Routes, manages core logic such as user authentication, adaptive learning-path control based on learner achievement, scoring-data processing, and the enforcement of a mandatory reflection interval between the first and second scoring rounds.

The RAG engine layer, the technical core of the system, consists of three tightly coupled modules. The query processor reconstructs a learner's input—whether a free-form question during criterion study or a scoring result awaiting feedback—into an optimized retrieval query. The retriever performs hybrid retrieval that combines BM25 keyword matching with dense-vector similarity, using the BAAI/bge-m3 model deployed on the institutional server for embeddings. By jointly considering the specific rubric terms captured by BM25 and the contextual meaning of the query captured by the dense vectors, this hybrid approach yields a retrieval precision that neither method alone could achieve. The retrieved documents—rubric annotations, expert scoring rationales, anchor papers, and similar past cases—are passed to the generator and synthesized into consistent, context-grounded feedback.

The generator employs a primary-fallback model strategy to ensure service continuity. Under normal operation, EXAONE 3.5 7.8B serves as the primary model, and if the primary model fails to respond within a specified timeout or produces an error, the system automatically switches to SOLAR 10.7B. Because both models were developed by Korean companies and share similar characteristics in maintaining a formal register and handling educational terminology, the design minimizes the effect of switching on the learner's experience. The effect of differences in output characteristics between the models on feedback quality, however, requires verification in a future implementation phase. As a final safeguard, the system can provide responses cached in the vector database for similar queries.

The database layer integrates PostgreSQL 15 with the pgvector extension to manage relational and vector data within a single system. The vector data consist of expert scoring cases with detailed rationales and theoretical guidelines, embedded into 1,024 dimensions with bge-m3, and support the semantic similarity search of the RAG engine. The relational data comprise user profiles, scoring logs containing element-wise scores and rationales by round, and learning-progress records, supporting the adaptive learning path and the longitudinal tracking of learner development. This integrated structure ensures that all data reside on the institutional on-premise server, eliminating external data transmission and providing an environment in which sensitive student and teacher information is handled securely.

The 2-Round Learning Path and Seven-Stage Program

The program follows a 2-Round intervention model structured around a pretest–intervention–posttest logic. In Round 1, an initial scoring task measures a learner's baseline scoring ability; a sequence of instructional interventions—feedback, calibration, and expert-example study—then addresses the identified gaps; and in Round 2, the same responses are re-scored to measure improvement. This design allows a within-subject comparison of scoring performance before and after the learning intervention. Because the improvement observed in a design that re-scores the same responses may include a learning effect specific to those items and responses, however, the transferability of the improvement is verified through performance trends across subsequent sessions that handle different responses. Fig. 2 depicts the overall learning

path.

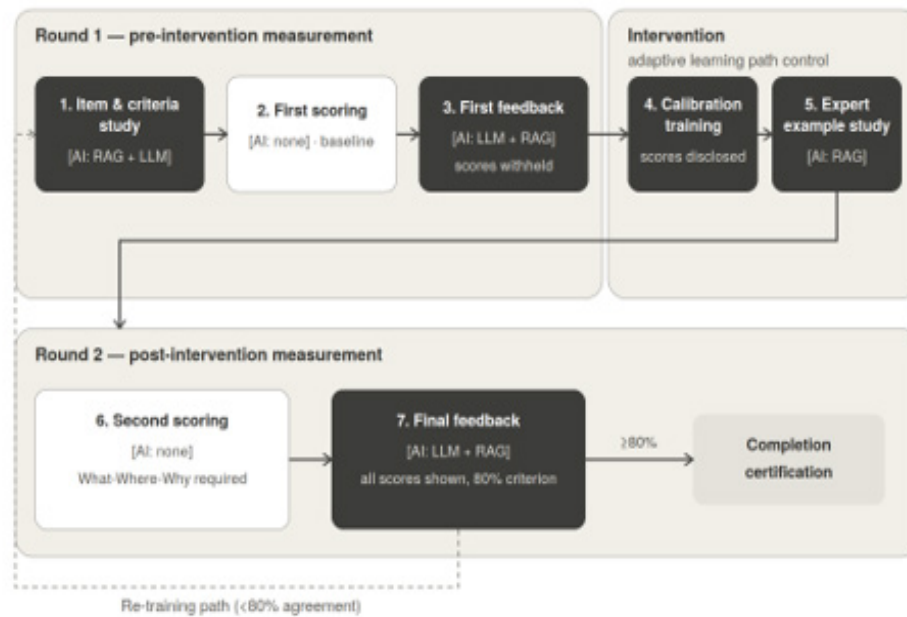


Fig. 2. 2-Round learning path

The seven stages, their learning objectives, and the AI role at each stage are summarized in Table 3.

Table 3. Seven-stage program overview

Stage	Name	Learning objective	AI role	Key activity
1	Item & criteria study	Develop thorough understanding of the assessment item and analytic rubric	RAG + LLM	Learners study the item and rubric; RAG retrieves explanations and expert rationales in response to free-text queries
2	First scoring	Measure pre-intervention scoring ability	None	Learners score one student response by rubric element; no rationale writing required
3	First feedback	Identify scoring gaps and provide improvement direction	LLM + RAG	Instant feedback (key improvement point, overall agreement rate) and detailed element-by-element analysis; expert scores withheld
4	Calibration training	Correct scoring tendencies and strengthen weak elements	LLM + Statistics + RAG	Learners score three anchor papers selected for their weak elements; leniency/severity tendency analyzed and reported
5	Expert example study	Internalize criterion application through expert cases	RAG	Three same-score responses and one to two borderline cases retrieved with full expert rationales
6	Second scoring	Measure post-intervention scoring ability	None	Same response re-scored; written scoring rationale required for every element
7	Final feedback	Evaluate growth and determine achievement	LLM + RAG	Round 1 vs. Round 2 comparison; agreement change quantified; achievement judged against 80% criterion

This sequence embodies a pedagogical rationale. The two scoring stages, Stages 2 and 6, deliberately exclude all AI intervention so that the scores recorded in each round reflect the learner's independent judgment rather than AI assistance, thereby preserving the validity of the pretest-posttest comparison. Stages 3 through 5, between these two

measurement points, constitute the instructional core of the program. Stage 3, the first feedback stage, reveals where a learner's scoring diverges from expert consensus without disclosing the experts' specific scores, focusing attention on understanding the criteria rather than memorizing target numbers. Stage 4, the calibration training stage, narrows the gap by providing anchor papers selected to match the learner's weak elements and scoring tendencies, with the expert scores now disclosed to enable explicit criterion alignment. Stage 5, the expert-example study stage, deepens understanding by presenting responses that received the same score but differ in their characteristics, along with borderline cases that required careful deliberation even among experts.

A key design decision lies in the strategic disclosure and withholding of expert scores by stage. In Stage 3, expert scores are kept undisclosed to prevent mechanical imitation and promote internalization of the criteria themselves; in Stage 4, the expert scores of the anchor papers are disclosed to enable direct calibration against a reference standard. In Stage 7, the first-round, second-round, and expert scores are all disclosed, allowing learners to survey their entire developmental trajectory. This gradual-disclosure strategy balances the competing goals of rater independence and criterion alignment.

The program does not determine training completion by a single score-agreement rate alone but uses two parallel indicators: first, the level of element-wise score agreement with expert scoring, and second, the quality of the scoring rationale written in the What-Where-Why structure. On the score side, 80% element-wise agreement is set as a provisional threshold, but this serves only as an indicator of convergence at the score level and does not substitute for the validity of the rationale; that is, a learner must both reach scores similar to the experts' and be able to validly explain the criterion-application process from which those scores were derived. This threshold remains open to adjustment based on field-application results. Learners who meet the criteria receive completion certification for the relevant item set, while those who do not are guided to a retraining path focused on their remaining weaknesses. Each session addresses one student response in the seven-stage sequence, and teachers are expected to complete multiple sessions to develop consistent scoring ability across responses of varying performance levels. In consideration of the time burden of completing all seven stages, however, the program is designed to be operated in a basic version centered on the essential procedures and an extended version that includes additional scoring of anchor papers, borderline-case analysis, and in-depth comparison of expert rationales, providing differentiated training paths according to a learner's scoring experience and proficiency.

Stage-Specific Roles of RAG and LLM

The relative weight of RAG and the LLM varies by stage according to the nature of the task. At stages where the accurate citation of expert cases and rubric grounds is paramount, the weight of retrieval is greater, whereas at stages requiring analysis of the causes of score gaps or the generation of individualized feedback, the weight of language-model reasoning is greater. The AI-role configuration of each stage is summarized in the AI role column of Table 3.

A distinctive element of the program is an adaptive feedback mechanism that adjusts both the content and the tone of feedback according to a learner's diagnosed scoring tendency. In Stages 3 and 4, the system computes the element-wise deviation between the learner's scores and the expert consensus scores; a tendency diagnosis based on the few cases of

a single session is treated as a provisional estimate and is refined as sessions accumulate. When the aggregate deviation indicates a leniency tendency, the LLM is prompted to emphasize analytic rigor, directing the learner's attention to under-applied criteria and presenting expert cases in which similar responses received lower scores with detailed justification. Conversely, when a severity tendency is detected, the LLM shifts to a developmental perspective, highlighting the strengths of a student response that the learner may have overlooked and presenting expert cases in which similar responses were scored more leniently. This tendency-responsive adjustment ensures that learners receive feedback tailored to their specific bias pattern rather than generic corrective information.

The system prompt that governs all LLM-generated feedback embeds a consistent persona: a peer assessor who guides the learner toward expert consensus rather than an authority who imposes correct answers. This framing is maintained across all feedback in Stages 3, 4, and 7, enhancing psychological acceptability and reinforcing the program's orientation toward emphasizing criterion understanding over score agreement.

Scoring Rationale Writing and Metacognitive Reflection

A key difference between the two scoring rounds is that in Stage 6, the second scoring round, learners must write a structured scoring rationale for every rubric element, whereas in Stage 2, the first scoring round, they do not. This requirement serves two purposes: it externalizes a learner's scoring thought process to promote metacognitive reflection, and it provides the RAG system with analyzable input for generating tailored final feedback.

The scoring rationale follows the What-Where-Why structure, summarized in Table 4.

Table 4. What-Where-Why scoring rationale structure

Component	Description	Well-formed example	Poorly formed example
What (Rubric citation)	Identify the specific rubric descriptor applied to the scoring decision	"[Content, 4 points] Rubric: 'Clear claim supported by two or more specific pieces of evidence'"	(No rubric reference)
Where (Textual evidence)	Cite concrete evidence from the student response that supports the judgment	"Evidence: citation of research findings (paragraph 2); observational example (paragraph 3)"	"The content was good" (no evidence cited)
Why (Judgment explanation)	Explain the logical reasoning connecting the rubric descriptor to the cited evidence	"The claim is clearly articulated and two pieces of evidence—one research-based, one experiential—provide convincing support, meeting the descriptor for level 4"	"It seemed good" (subjective impression without reasoning)

The What element anchors the scoring rationale to the formal language of the rubric, enabling learners to identify which specific descriptor applies rather than relying on a general impression. The Where element requires learners to locate a specific passage in the student response, so that scoring decisions rest on textual evidence rather than holistic intuition. The Why element requires learners to articulate the logical connection between the cited rubric descriptor and the identified evidence, making explicit and reviewable the most tacit reasoning step in expert scoring.

This three-element structure is designed to promote metacognitive reflection through externalization. In expert scoring, the mapping from rubric to evidence to judgment typically occurs as a rapid and largely automated cognitive process (Collins et al., 1989). Novice raters, however, benefit from decomposing this process into discrete, verbalizable steps,

which corresponds to the articulation phase of Cognitive Apprenticeship, in which learners are asked to make their tacit knowledge explicit. By requiring the scoring rationale only in Stage 6 and not in Stage 2, the program exploits this contrast: a learner who scored intuitively in Stage 2 must justify every decision in Stage 6, which heightens awareness of how their scoring criteria have changed—or persisted—through the intervening training.

To prevent novice raters from failing to externalize their scoring thought process because of the burden of writing rationales, the system provides UI-level scaffolding during the writing process. The rationale input fields on the second scoring screen are structured into the three fields of What, Where, and Why, with the sentence frames of Table 5 presented as guiding prompts in each field. Learners can consult one or two model rationales through a 'view examples' function, which provides examples only from items other than the one currently being scored to prevent mechanical imitation of phrasing. While writing, the system checks in real time whether all three elements are present and whether the rubric citations are accurate, prompting the learner about any missing element. This scaffolding is designed to be gradually removed as a learner's proficiency improves, ultimately aiming at independent rationale writing without external support.

Table 5. Sentence frames for scoring rationale writing

Element	Sentence frame (input-field prompt)	Example
What	Applying the '[descriptor]' criterion of rubric element '[element name]'	Applying the 'clear claim and two or more specific pieces of evidence' criterion of rubric element 'Content (4 pts)'
Where	The '[relevant passage]' at [location: paragraph/sentence] in the student response corresponds to this	The research-finding citation in paragraph 2 and the observational example in paragraph 3 of the student response correspond to this
Why	Because [evidence] meets (or fails to meet) [the key requirement of the descriptor], I judge it to be [score] points	Because the claim is clear and two research- and experience-based pieces of evidence are presented convincingly, meeting the requirement for the 4-point level, I judge it to be 4 points

The RAG system analyzes a submitted rationale along two dimensions. On the formal dimension, it checks whether all three elements are present and whether the rubric descriptors are accurately cited; on the content dimension, it computes the semantic similarity between the learner's rationale and the corresponding expert rationale stored in the vector database. When a gap is detected—such as a missing evidence citation or reasoning that substantially diverges from the expert logic—the system retrieves a model rationale from the database and presents it as a comparative reference, allowing the learner to see how an expert would have justified the same scoring decision. This rationale-level feedback is provided in Stage 7, the final feedback stage, complementing the score-level feedback of the earlier stages. Whereas Stages 3 and 4 address what score a learner assigned and how it differs from the experts', Stage 7 addresses why the learner arrived at that score and how the underlying reasoning compares with expert judgment.

By thus integrating structured rationale writing into the 2-Round structure, the program makes the cognitive process of scoring itself a target of training and feedback.

Technical Feasibility Demonstration

To verify the technical feasibility of the proposed RAG pipeline, a proof-of-concept was implemented with the same configuration described in this paper. EXAONE 3.5 7.8B (4-bit quantization, Q4_K_M) was used as the generation

model and bge-m3 (1,024 dimensions) as the embedding model, and retrieval was configured as a hybrid method combining BM25 keyword search and dense-vector search via Reciprocal Rank Fusion (RRF). All components were run locally in a single-GPU environment without external APIs, confirming consistency with the on-premise deployment assumption. A small Korean-language case set simulating rubric annotations, anchor papers, and expert scoring rationales was used as the knowledge database. Fig 3(a) shows the runtime data flow in which a learner's scoring input is transformed into a query, undergoes hybrid retrieval and evidence-grounded generation, and is output as feedback.

Fig 3(b) illustrates the adaptive feedback mechanism with two cases, showing that the same pipeline retrieves different expert materials according to a learner's scoring tendency. In the leniency-tendency case, where the learner assigned 4 points and the expert 3, leniency-correction rationales and a single-evidence case were retrieved at the top; in the severity-tendency case, where the learner assigned 2 points and the expert 3, severity-correction rationales and an expression-error case were retrieved at the top; and the generated feedback was differentiated to match the two tendencies based on the retrieved evidence. This shows that the same system can adaptively compose feedback according to the direction of scoring bias.

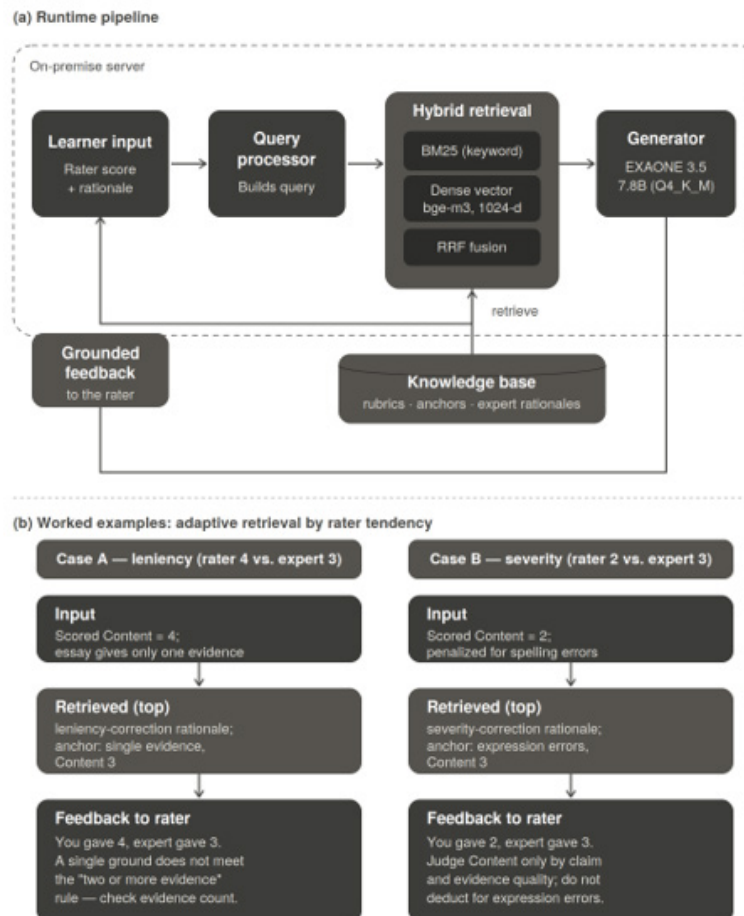


Fig. 3. RAG runtime pipeline and worked examples of adaptive feedback

This demonstration, however, is a confirmation of technical feasibility for a limited set of cases using a small knowledge database; it is distinct from a systematic evaluation of feedback-generation quality or a verification of the educational effectiveness of the adaptive feedback, which are tasks for future research.

Expert Review of the Program Design

To formatively review the validity of the designed program, an expert review was conducted with seven in-service Korean Language Arts teachers. The panel consisted of secondary Korean Language Arts teachers with experience in essay-type assessment, and prior to the review they were provided with an overview document describing the program's design principles, learning path, stage composition, and methods for constructing the materials and knowledge database.

The review instrument was a survey composed of six areas—the validity of the design principles, the appropriateness of the program structure and operation, field applicability, the suitability of the assessment materials and knowledge-database construction, and overall open-ended comments—with each item rated on a four-point scale (1 = strongly disagree, 4 = strongly agree). Responses were analyzed using the item-level content validity index (I-CVI) and the scale-level content validity index (S-CVI/Ave), with an I-CVI of .78 or higher (endorsement by at least six of the seven panelists) set as the criterion for establishing validity (Lynn, 1986; Polit & Beck, 2006). A chance-corrected modified kappa (k^* ; Polit, Beck, & Owen, 2007) was reported alongside, and the open-ended responses were summarized through thematic analysis.

The review found that 23 of the 24 rated items met the criterion of I-CVI .78 or higher, and S-CVI/Ave was .97. The chance-corrected modified kappa (k^*) was also .85 or higher—an 'excellent' level—for 23 items, with the remaining item at .66, a 'good' level. The mean I-CVI exceeded .90 in all four areas (Table 6), confirming strong content validity. The panel comprised three teachers with fewer than 5 years of experience, three with 5–10 years, and one with 11–20 years; five worked at middle schools and two at high schools; and their scoring and training experience was diversely distributed.

Table 6. Content validity results by domain

Domain	Items	I-CVI range	Items $\geq$.78	Mean I-CVI
Design principles (B)	7	1.00	7 / 7	1.00
Program structure (C)	8	.86 – 1.00	8 / 8	.98
Field applicability (D)	5	.71 – 1.00	4 / 5	.91
Data & knowledge DB (E)	4	.86 – 1.00	4 / 4	.96
Overall	24	.71 – 1.00	23 / 24	S-CVI/Ave = .97

The single item that fell below the criterion was the field-applicability item 'It is realistically feasible for a teacher to complete all seven stages while carrying out school duties' (I-CVI = .71), which reflects a reservation about practical feasibility rather than about the content validity of the design itself. The three remaining items that did not reach unanimity concerned the 80% element-wise completion criterion, the web-browser access method, and the selection criterion for inter-rater agreement cases (each I-CVI = .86), and were examined together with the open-ended comments.

In the open-ended responses, reviewers identified the following strengths of the design: first, that it treats rater training

not as a process of simply matching expert scores but as one in which learners externalize and examine their own grounds for judgment in the What-Where-Why form; second, that its systematic multi-stage structure—first scoring, feedback, calibration, second scoring, and final feedback—distinguishes it from one-off training; and third, that its non-face-to-face environment and use of RAG-based expert knowledge mitigate the spatiotemporal constraints of face-to-face rater deliberation.

The reviewers' principal reservations and suggestions, and the corresponding design refinements, are as follows. First, many reviewers identified the volume and complexity of the seven stages, and the burden of carrying them out alongside school duties, as the greatest obstacle to field adoption. In response, the tiered basic and extended operating scheme described above was introduced, differentiating training paths by proficiency. Second, one reviewer noted the need to distinguish whether the 80% element-wise agreement criterion refers to agreement on quantitative scores or on the score-derivation process. Accordingly, it was clarified that the 80% agreement is limited to an indicator of convergence at the score level and that rationale quality is included in parallel in the completion judgment. Third, several reviewers noted the need to examine whether the program generalizes beyond a specific essay task and rubric to diverse essay-type assessment situations. This study acknowledges this as a limitation and sets the refinement of task-specific criteria and extension to diverse tasks as future tasks. Fourth, there were suggestions regarding the transparency of the expert scoring-and-consensus procedure and the use of borderline cases in which expert judgments diverge. The design already includes borderline-case designation in the knowledge-database construction procedure, and reflecting this suggestion, it makes the two raters' cross-review and consensus procedure and its outputs visible to learners, transparently presenting the process by which expert judgments are derived. Finally, one reviewer raised the fundamental question of how the convergence of rater training toward 'consistency' reconciles with the assessor's discernment that is the essence of essay-type assessment; the program seeks to partially ease this tension by pursuing both the securing of score consistency and the qualitative refinement of scoring rationales.

This expert review is a formative evaluation of the content validity of the design and is distinct from an empirical verification of the program's educational effectiveness. Moreover, because it is based on a single round and a small panel, its results should be complemented through subsequent implementation and effectiveness studies.

Conclusions and Implications

Summary

This study designed a RAG-based rater training program for Korean Language Arts essay assessment. Drawing on Cognitive Apprenticeship, Frame-of-Reference training, and Deliberate Practice theory, six core design principles were derived and implemented as a seven-stage program following a 2-Round learning path. Built on open-source Korean-language LLMs—EXAONE 3.5 7.8B and SOLAR 10.7B deployed on an institutional on-premise server—the program's RAG system differentiates the roles of retrieval and generation across stages to provide personalized, context-grounded feedback, and promotes metacognitive reflection through a structured What-Where-Why scoring rationale requirement.

Implications

This design resolves, within a single integrated architecture, the dual limitation identified in the Introduction: the spatiotemporal constraints of face-to-face training and the superficial feedback of existing computer-based systems. Existing online rater training has been limited to indicating whether a learner's scores match a benchmark (Elder et al., 2007) or to providing statistical summaries of scoring accuracy (Choi, Y. et al., 2019), failing to explain why a particular scoring decision diverges from expert consensus. The RAG-based feedback mechanism of this study moves beyond score-level comparison by retrieving contextually relevant expert rationales and generating individualized explanations based on a learner's scoring patterns, and the What-Where-Why rationale analysis in Stage 7 extends this to the reasoning process, making the cognitive dimension of scoring a direct target of training and feedback.

From a theoretical perspective, this study demonstrates that the principles of Cognitive Apprenticeship (Collins et al., 1989) can be implemented in a digital environment without losing their pedagogical intent. The six instructional methods—modeling, coaching, scaffolding, articulation, reflection, and exploration—map onto the program's seven stages. Expert rationales and anchor papers perform modeling; RAG-generated feedback performs coaching and scaffolding; the What-Where-Why rationale requirement performs articulation; and gradual disclosure and the developmental feedback of Stage 7 perform reflection, while the learner's independent second scoring corresponds to the independent performance that the exploration phase targets. This suggests that RAG technology can serve as a medium for transforming apprenticeship-based pedagogy into scalable, self-directed expertise development—a proposition that requires empirical verification in future research.

Furthermore, redefining RAG as a teacher-training tool rather than a student-facing information system carries implications for the field of educational AI as a whole. To date, educational applications of RAG have mostly concentrated on learner-facing Q&A or automated scoring; by repositioning RAG as a mechanism for developing the professional judgment of human assessors, this study presents a complementary line of application in which AI strengthens rather than replaces humans in assessment.

These implications can be summarized as three contributions. First, instead of applying technology post hoc, the study derived design principles from pedagogical theory and identified RAG as the technology to implement them, thereby presenting a design methodology in which educational goals drive technical choices. Second, it provides a framework for teacher professional development suited to the Korean educational context in the area of assessment expertise, where systematic, practice-based tools have been particularly scarce. Third, the on-premise deployment strategy using open-source quantized LLMs demonstrates that a privacy-preserving, institution-controlled AI training system is feasible without reliance on external cloud APIs, offering a practical model for educational institutions under strict data-protection requirements.

Limitations and Future Research

This study has three main limitations. First, as a design-stage study, it has not undergone the cycle of implementation, field application, evaluation, and revision; the design principles and architecture presented here are therefore proposals grounded in theoretical rationale, and their educational effectiveness must be empirically verified in future research.

Second, the assessment data used by the program are general writing-assessment data centered on argumentative writing and do not cover the full range of Korean Language Arts essay assessment, such as descriptive assessment in the reading and literature domains; a phased expansion of the knowledge database with domain-specific items and scoring cases is therefore needed. Third, because the program is currently limited to Korean Language Arts, generalizing it to other subjects such as mathematics, science, and English, or to other assessment types such as performance assessment and portfolio assessment, would require subject-specific adaptation of the rubric database and feedback design.

Two directions for future research follow from these limitations. First, implementing the program, deploying it to in-service teachers, and conducting a quasi-experimental effectiveness study comparing RAG-based training with traditional face-to-face calibration would complete the next cycle of design-based research. Second, the program could be extended to other subjects and school levels, building curriculum-based knowledge databases suited to each context.

Finally, two recommendations are offered. Future research should evaluate LLMs not only with general-purpose benchmarks but also with education-domain-specific benchmarks, because the requirements for generating pedagogically appropriate rater feedback may differ from what standard NLP evaluation captures. In addition, although the current design supports an individual teacher's independent completion of the seven stages, it should in the future incorporate community-of-practice-based collaborative features such as inter-rater discussion boards, shared rationale repositories, and peer feedback. The web-based on-premise architecture already provides the technical foundation for such an extension, and for large-scale deployment, domestic cloud services could be used to scale beyond a single institutional server while maintaining within-border data residency.

Acknowledgements

This work was supported by the Master's Research Encouragement Grant funded by the National Research Foundation of Korea (NRF-2025S1A5B5A20020721).

This article is based on a revised version of the author's master's thesis, A Study on Developing a Rater Training Program for Korean Language Arts Essay Assessment Based on Retrieval-Augmented Generation (RAG), submitted to the Graduate School of Korea National University of Education in 2026.

Conflicts of Interest

The author declared no conflicts of interest.

References

- Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, 6(2), 205–212. <https://doi.org/10.5465/amr.1981.4287782>
- Choi, Y., Kim, S., Lee, S., & Park, Y. (2019). Development a Python-based writing assessment practice program. *Journal of Learner-Centered Curriculum and Instruction*, 19(12), 769–787. <https://doi.org/10.22251/jlcci.2019.19.12.769>

- Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Hillsdale, NJ: Erlbaum.
- Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 24(1), 37–64. <https://doi.org/10.1177/0265532207071511>
- Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 683–704). Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/CBO9780511816796.038>
- Jang, S. Y. (2010). The development and evaluation of a systematic training program for increasing both rater reliability and rating accuracy (Doctoral dissertation). University of Illinois at Urbana-Champaign, Champaign, IL.
- Kim, D., Park, C., Kim, S., Lee, W., Song, W., Kim, Y., Kim, H., Kim, Y., Lee, H., Kim, J., Ahn, C., Yang, S., Lee, S., Park, H., Gim, G., Cha, M., Lee, H., & Kim, S. (2024). SOLAR 10.7B: Scaling large language models with simple yet effective depth up-scaling. arXiv preprint arXiv:2312.15166. <https://doi.org/10.48550/arXiv.2312.15166>
- Lee, Y. (2015). The effects of writing assessment workshop on the writing assessment expertise of pre-service writing teachers. *Research on Writing*, 27, 83–117.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). <https://doi.org/10.48550/arXiv.2005.11401>
- LG AI Research. (2024). EXAONE 3.5: Series of large language models for real-world use cases. arXiv. <https://doi.org/10.48550/arXiv.2412.04862>
- Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382–386. <https://doi.org/10.1097/00006199-198611000-00017>
- Ministry of Education. (2022). 2022 revised Korean Language Arts curriculum (Notification No. 2022-33). Ministry of Education, Republic of Korea.
- Ministry of Education. (2023, October 10). Finalized plan for the 2028 college admissions reform [Press release]. Ministry of Education, Republic of Korea.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189–227.
- National Information Society Agency. (2021). 에세이 글 평가 데이터 [Essay writing assessment data] [Data set]. AI-Hub. Retrieved June 20, 2026, from <https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=545>
- OECD. (2018). *The future of education and skills: Education 2030*. Paris: OECD Publishing. <https://doi.org/10.1787/54ac7020-en>
- Park, J. (2013). A study on development of the on-line rater training program for writing assessment. *Research on Writing*, 17, 413–440.
- Park, J. (2014). A study on the cognitive scoring process model in the writing assessment. *Cheongnam Korean Language Education*, 52, 57–85.
- Park, Y. (2011). Promotion specialities of Korean language teachers in writing assessment. *Research on Writing*, 13, 89–116.

- Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29(5), 489–497. <https://doi.org/10.1002/nur.20147>
- Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, 30(4), 459–467. <https://doi.org/10.1002/nur.20199>
- Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55(3), 2495–2527. <https://doi.org/10.1007/s10462-021-10068-2>
- Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, 72(7), 534–539. <https://doi.org/10.2307/20404455>
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263–287. <https://doi.org/10.1177/026553229801500205>

Author's Information

Son, Jin-Hwan: Korea National University of Education, Graduate Student, First Author. <https://orcid.org/0009-0000-4984-2725>