

DOI: 10.37943/LSGM5046

Syrym Zhakypbekov

Master's Degree, Assistant Professor, Department of Computer Engineering
s.zhakypbekov@iitu.edu.kz, orcid.org/0000-0001-9112-5922
International Information Technology University, Kazakhstan

Artem Bykov

PhD, Associate Professor, Department of Information Systems
a.bykov@iitu.edu.kz, orcid.org/0000-0002-9563-5185
International Information Technology University, Kazakhstan

Vera Yermakova

Candidate of Philological Sciences, Associate Professor, Department of Languages
v.yermakova@iitu.edu.kz, orcid.org/0009-0004-7347-1580
International Information Technology University, Kazakhstan

Regina Sharshova

Master's Degree, Assistant Professor, Department of Languages
r.sharshova@iitu.edu.kz, orcid.org/0000-0002-7367-6007
International Information Technology University, Kazakhstan

**CRITERION-GUIDED AI ARCHITECTURE FOR LEARNING SUPPORT AND
PRELIMINARY STUDENT ASSESSMENT**

Abstract: This paper presents a pilot evaluation of a criterion-guided human-in-the-loop architecture for an intelligent educational system that supports learning and preliminary assessment of students' written and oral responses. Instead of assigning a final score directly, the system combines relevant context, formalized criteria, interpretable response features, supporting evidence, uncertainty estimates, and instructor verification. The architecture separates probabilistic components, including retrieval, automatic speech recognition, natural language processing, and a language model, from deterministic assessment and governance mechanisms such as rubric weights, aggregation rules, thresholds, audit logging, and referral rules. Two modes were implemented: a retrieval-augmented generation assistant for formative learning support and an AI examiner for preliminary criterion-based assessment. In the assistant pilot, 336 of 347 queries were processed without recorded technical errors, giving a completion rate of 96.8%; source citations were included in 89.0% of responses, and median response time was 980 ms. Student perceptions of the examiner were evaluated using 52 questionnaires on a five-point Likert scale. Mean ratings were highest for question clarity at 4.81, usefulness of criterion-level feedback at 4.46, and sufficiency of response time at 4.42. Students also reported notable uncertainty about the grade, with a mean rating of 3.79, and concern about speech-recognition errors, with a mean rating of 3.31. Among 45 oral responses, 12 (26.7%) lacked usable word-level timestamps and were referred for review. A worked case demonstrated traceability from response features and evidence to criterion scores, uncertainty handling, aggregation, and instructor verification. A preliminary comparison of 23 paired AI and instructor-released scores was treated as diagnostic because of a severe ceiling effect. The results support technical feasibility and procedural acceptability, but not yet the validity or reliability of automated assessment.

Keywords: intelligent education systems; retrieval-augmented generation; rubric-based assessment; explainable artificial intelligence; human-in-the-loop; uncertainty estimation; higher education; decision support.

Introduction

The use of artificial intelligence in higher education is increasingly extending beyond information and advisory services to personalized learning, explanations of course content, feedback generation, analysis of student difficulties, and assessment of learning outcomes [1].

Advances in generative language models have broadened the capabilities of intelligent educational systems by enabling them to process natural-language queries, analyze extended responses, and produce context-sensitive explanations. Their adoption, however, requires attention not only to functionality but also to the reliability, transparency, and pedagogical defensibility of the conclusions they generate.

Two functionally distinct uses of artificial intelligence can be identified in education. The first is formative learning support, in which a system helps a student locate information, obtain an explanation, or identify gaps in understanding. The second is response analysis and assessment, where the system's output may influence the evaluation of learning outcomes and the student's subsequent academic trajectory. An inaccurate answer from a learning assistant can be clarified through dialogue; an error in automated assessment, by contrast, calls for much stricter control of validity, fairness, and the evidential basis of the decision [2-3].

Formalized criteria, supporting evidence, uncertainty estimation, and human oversight are particularly important in assessment settings. A language model should not serve as an autonomous source of the final score. It is better treated as one component of a broader computational pipeline, assigned bounded tasks such as text processing, feature extraction, criterion matching, and preparation of a preliminary, interpretable judgment. The final academic decision should remain with the instructor, in keeping with human-in-the-loop principles and a risk-based approach to artificial intelligence [4-5].

This study proposes a criterion-guided architecture for an intelligent educational system operating in two modes: a RAG assistant for formative learning support and an AI examiner for preliminary, criterion-based analysis of students' written and oral responses. The modes share several core mechanisms: retrieval of relevant context and supporting passages, application of formalized criteria, estimation of reliability and uncertainty, logging, and instructor verification. This design frames learning support and preliminary assessment not as isolated services, but as complementary modes within a single governed intelligent system.

Literature Review and Problem Statement

Intelligent educational systems have traditionally incorporated mechanisms for representing domain knowledge, modelling learner state, selecting pedagogical interventions, and managing user interaction. Recent research has expanded this structure to include natural-language processing, personalized feedback, learning-activity analytics, and generative models [1, 6]. Such systems are now used both to support independent learning and to automate selected instructional tasks, including response analysis and the preparation of assessment feedback.

Retrieval-augmented generation is one of the more recent approaches to formative learning support. In RAG systems, a language model generates its response using passages retrieved from an external knowledge base [7-8]. In an educational setting, this approach can restrict the model's context to course-specific materials and provide students with references to the sources used. Yet the presence of relevant documents in the context does not guarantee that every claim in the generated response is supported by those documents. Research on groundedness shows that language models may still produce unsupported claims even when the retrieved context is appropriate [9]. An educational RAG assistant must therefore monitor both retrieval quality and the extent to which substantive claims are anchored in evidence from course materials.

A second strand of research concerns the automated assessment of written and oral responses. Systematic reviews indicate that artificial intelligence can be applied to essays, short constructed responses, source code, and other learning products [2, 10]. Contemporary systems based on large language models can account for response content, structure, argumentation, and adherence to task instructions. Their scores, however, remain sensitive to task wording, model choice, prompt design, subject area, and response language [11]. Agreement between a system's overall score and an instructor's score is therefore not, by itself, sufficient evidence of assessment validity.

Analytic rubrics offer one route to greater transparency in automated assessment. By decomposing performance into separate criteria, a rubric supports both an overall score and criterion-level feedback [12]. Under an evidence-centered approach, an assessment decision should rest on an explicit relationship among the target construct, the task, observable response features, and the evidence supporting the inference [13]. For an intelligent system, this means retaining, for every criterion, its description, the derived features, the criterion score, the supporting passages, and an explanation of how the decision was reached.

A further body of work addresses the reliability and control of automated inferences. Uncertainty estimation helps identify cases in which a model's decision should not be used without additional review [14]. Selective classification allows a system to abstain when confidence is insufficient [15]. In educational assessment, abstention becomes a referral mechanism: a response should be passed to an instructor when transcription quality is poor, evidence is inadequate, uncertainty is high, or criterion-level results conflict. Human oversight is also required by current ethical principles for AI in education and by risk-based requirements governing systems that affect the assessment of learning outcomes [3].

The literature reviewed above generally treats RAG assistants, automated assessment, uncertainty estimation, and human oversight as separate lines of inquiry. RAG research tends to emphasize retrieval and the generation of grounded responses, whereas automated-assessment studies focus on score agreement, feedback quality, and consistency with expert judgments. Logging, evidential traceability, and referral to a human reviewer are more often treated as supplementary controls than as a shared architectural layer of an educational system.

What remains insufficiently formalized is a unified computational pipeline in which learning support and preliminary assessment share mechanisms for criterion-based analysis, evidential grounding, uncertainty estimation, auditing, and instructor verification. For every criterion, the system should provide a traceable link among the input data, derived features, supporting passages, criterion score, reliability indicator, and the reason for referral to an instructor. Addressing this gap requires a criterion-guided human-in-the-loop architecture that separates probabilistic AI components from deterministic rules governing the production, control, and approval of preliminary decisions.

The aim of this study is to develop and pilot-evaluate a criterion-guided human-in-the-loop architecture that combines RAG-based learning support with preliminary analysis of students' written and oral responses using a formalized assessment rubric.

The study addresses the following objectives:

- develop a shared architecture for an intelligent educational system that integrates the RAG-assistant and AI-examiner modes;
- define the probabilistic and deterministic layers used to process data and generate preliminary decisions;
- formalize the relationship among rubric criteria, derived response features, supporting evidence, and criterion-level scores;
- develop a model for aggregating criterion-level scores that accounts for normative criterion weights and the reliability of the results;
- develop a mechanism for uncertainty estimation and referral of insufficiently reliable decisions to an instructor;
- specify the technical and user-centered indicators used in the pilot evaluation of system feasibility and perception; and
- pilot the RAG assistant and evaluate students' perceptions of the AI-examiner mode.

The contribution of this study is the formalization and pilot implementation of a governed human-in-the-loop assessment pipeline in which each preliminary decision can be traced from observable response features and supporting evidence to criterion-level scores, uncertainty handling, deterministic aggregation, and instructor verification. Architecture separates probabilistic AI components from deterministic assessment and governance rules and

operationalizes abstention and referral when the evidential basis is insufficient. The contribution is therefore architectural and methodological; the present pilot does not establish superior scoring accuracy or assessment validity, which require independent expert validation.

Methods and Materials

Overall Architecture of the Intelligent System

The proposed system follows a multilayer architecture comprising a user layer, an API layer, a data layer, an intelligent-processing layer, a criterion-analysis layer, a decision-explanation layer, and an expert-verification loop. This structure is consistent with modular accounts of contemporary intelligent tutoring systems, in which domain knowledge, learner data, pedagogical interaction mechanisms, intelligent-processing components, and the user interface are functionally separated [6]. The additional layers for criterion analysis, explanation, auditing, and instructor verification are aligned with human-centered AI architectures [16].

The user layer contains separate interfaces for students and instructors. The student interface supports question submission, access to explanations, task completion, and the submission of oral or written responses. The instructor interface supports the upload of course materials, rubric configuration, review of preliminary judgments, examination of disputed cases, and analysis of interaction logs.

The API layer handles request routing, session management, authorization, logging, language-model calls, and data exchange among subsystems.

The data layer comprises a relational database, a vector store, a repository of course materials, a database of verified responses, a rubric repository, model-call logs, and records of instructor verification.

The intelligent-processing layer contains the RAG module, automatic-transcription module, NLP-analysis module, feature-extraction module, LLM Gateway, structured-output module, and explanation-generation module.

The criterion layer maps input data to formalized criteria, calculates criterion-level indicators, aggregates the decision, estimates uncertainty, and determines whether instructor review is required.

The architecture supports two operating modes (Fig. 1): an AI assistant for formative learning support and an AI examiner for preliminary assessment. Both modes draw on shared components: a knowledge base, vector store, LLM Gateway, criterion module, uncertainty module, audit trail, and instructor-verification loop.

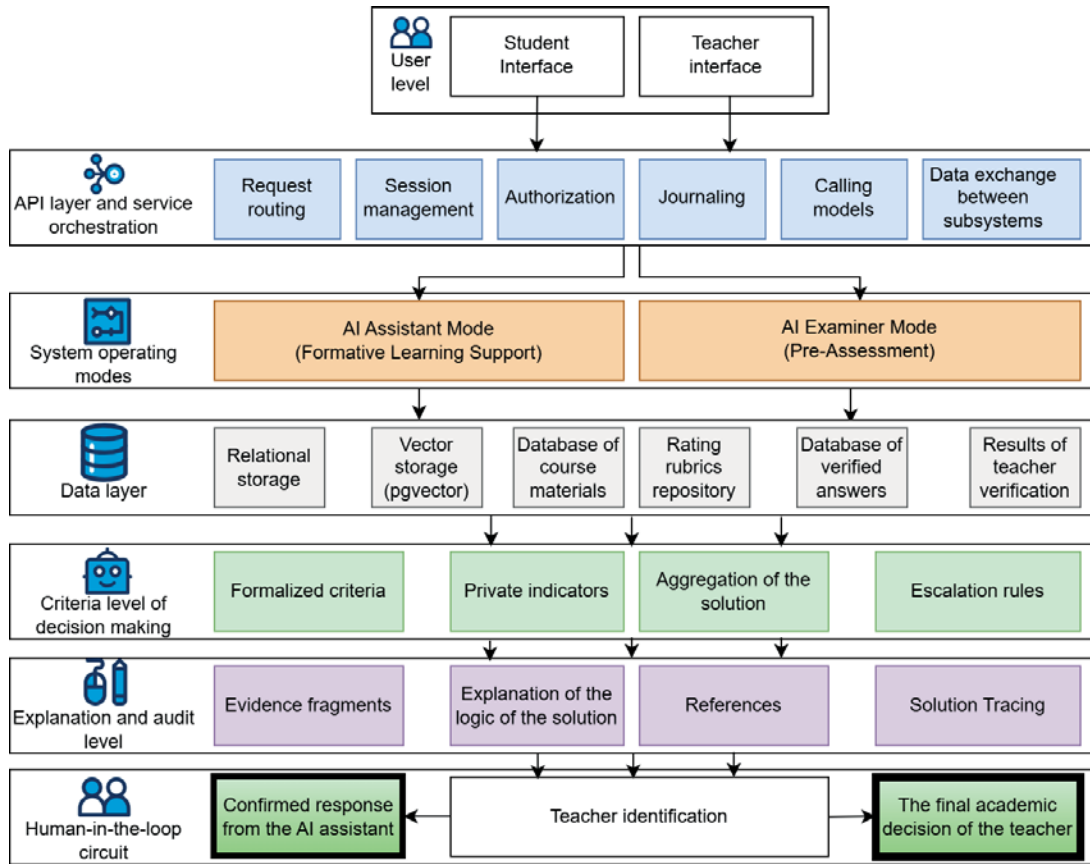


Figure 1. Overall architecture of the criterion-guided intelligent educational system

Technical Architecture of the AI-Assistant Mode

The AI-assistant mode is implemented as a RAG pipeline that generates responses from materials associated with a specific course. A user's query (q) first enters a preprocessing module, where the text is normalized, the language is identified, key terms are extracted, and a search representation is constructed. The query is then passed to a hybrid retrieval module that combines full-text and vector search.

The relevance of document (d) to query (q) is defined as follows:

$$Rel(q, d) = \lambda \cdot \cos(E_q, E_d) + (1 - \lambda) \cdot BM25(q, d), \quad (1)$$

where E_q and E_d are vector representations of the query and document, respectively; ($BM25(q, d)$) is the lexical relevance measure; and λ is the coefficient balancing semantic and full-text retrieval.

The set of retrieved passages is defined as follows:

$$R(q) = TopKd_j: Rel(q, d_j) \geq \tau_R. \quad (2)$$

If the relevance of the retrieved context falls below the specified threshold, the system should not produce a high-confidence answer. It should instead request clarification or refer the query to an instructor.

Accordingly, the following groundedness measure is introduced to assess whether the answer is supported by evidence:

$$G(a, R) = \frac{|Claims(a) \cap Supported(R)|}{|Claims(a)|}, \quad (3)$$

where $Claims(a)$ is the set of substantive claims in the answer and $Supported(R)$ is the set of claims supported by the retrieved passages from the course materials.

Figure 2 presents the technical workflow of the AI assistant. The user's query enters the hybrid retrieval module, after which relevant passages from the course materials are passed to the language model through the LLM Gateway. The generated response is checked for relevance, groundedness, and source attribution. When reliability is insufficient, the response is placed in the instructor-review queue.

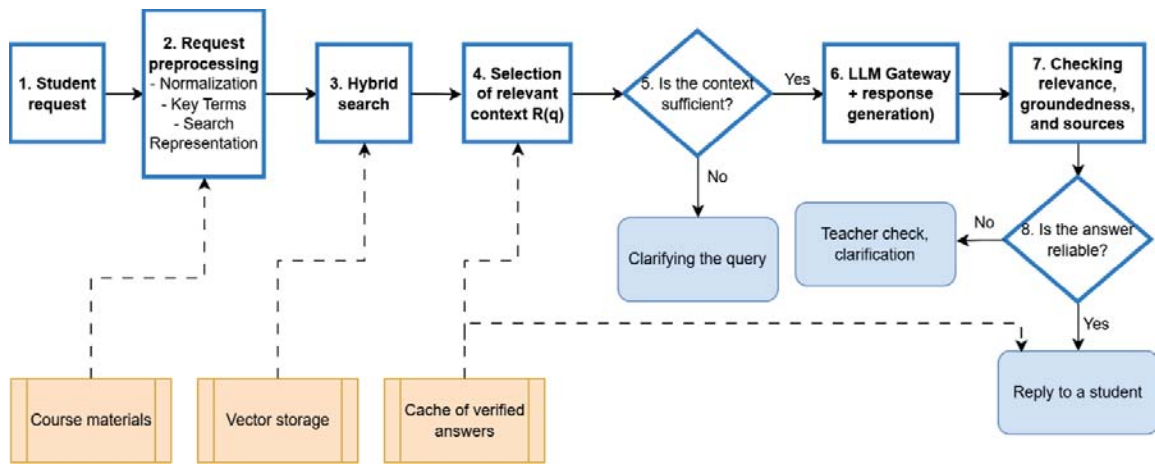


Figure 2. Technical workflow of the AI assistant's RAG mode

Technical Architecture of the AI-Examiner Mode

The AI-examiner mode is designed for preliminary, criterion-based analysis of a student's response. Its inputs are the examination task z , assessment rubric R_b , and student response x . If the response is delivered orally, it is first transcribed automatically:

$$x = ASR(audio). \quad (4)$$

Transcription quality is treated as a separate technical indicator because speech-recognition errors may distort subsequent analysis. When a reference transcript is available, quality can be evaluated using the Word Error Rate:

$$WER = \frac{S + D + I}{N}, \quad (5)$$

where S is the number of substitutions, D the number of deletions, I the number of insertions, and N the number of words in the reference transcript.

The quality of the speech-input pipeline is defined as follows:

$$Q_{ASR} = 1 - WER. \quad (6)$$

If $Q_{ASR} < \tau_{ASR}$, the system should not issue a high-confidence assessment judgment without manual review.

Following transcription, the response is segmented into meaning-bearing units:

$$x = x_1, x_2, \dots, x_l. \quad (7)$$

For each segment, the system derives features representing fluency, lexical diversity, grammatical accuracy, coherence, task coverage, and response structure.

Figure 3 presents the general architecture of the AI examiner. The examination task, rubric, and student response are provided as inputs. The system's core performs criterion-based NLP analysis, extracts response features, calculates a preliminary score accompanied by an uncertainty signal, and sends the result to the instructor for review.

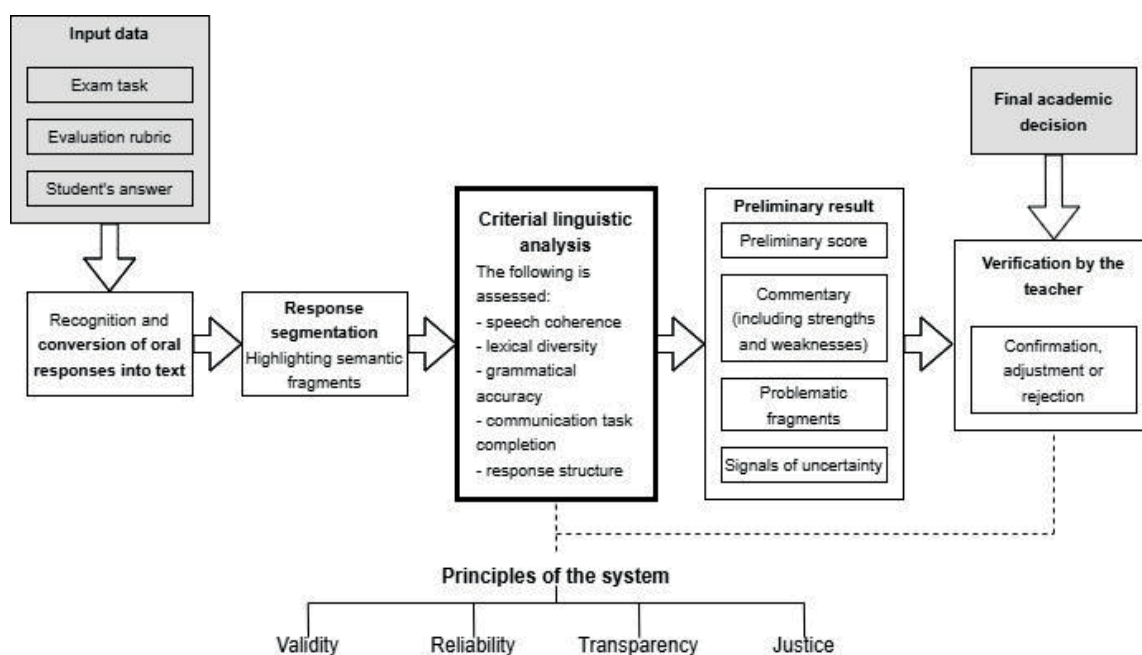


Figure 3. General architecture of the AI examiner

Extraction of Interpretable Response Features

Unlike holistic scoring, in which a model produces an overall score without making the basis of its decision explicit, the proposed system uses interpretable features tied to individual rubric criteria. This follows the logic of criterion-aligned automated assessment, which considers not only the response but also its structural, semantic, and criterion-relevant components [17]. Such decomposition is particularly important in information technology courses, where student responses may contain specialist terminology, English borrowings, professional jargon, and stable discourse patterns associated with IT professionals' linguistic worldview [18]. A comparable decomposition strategy has previously been used to analyze complex texts for emotion recognition [19]. Here, it is applied to the criterion-based analysis of student responses and to the justification of preliminary assessment decisions.

For an oral response, the system derives measures of speech fluency (rate and proportion of pauses), lexical diversity, grammatical accuracy, semantic relevance, and coherence.

Lexical diversity can be estimated using the type-token ratio:

$$TTR = \frac{N_{unique}}{N_{total}}. \quad (8)$$

The semantic relevance of a response to the task can be estimated from the cosine similarity between the response and task embeddings. This is consistent with the widely used sentence-embedding approach to semantic textual similarity [20]:

$$SemRel(x, z) = \cos(E_x, E_z). \quad (9)$$

Response coherence can be estimated as the mean semantic similarity between adjacent meaning-bearing segments. This approach is consistent with automated text-coherence research, in which local coherence is modelled through the interaction and semantic similarity of neighboring sentences or text segments [21].

Figure 4 shows the pipeline for extracting interpretable features from an oral response. The audio response is converted into a time-aligned transcript, which is then segmented and analyzed for fluency, lexical diversity, syntactic complexity, semantic relevance, and coherence. The resulting features are passed to the criterion-based assessment module.

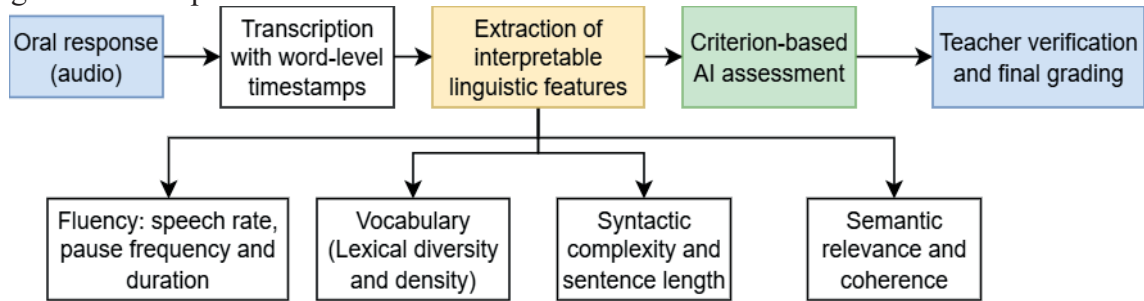


Figure 4. Transparent pipeline for extracting interpretable features from oral responses

Formalized Model of the Criterion-Based Decision

Let the set of criteria be $C = \{c_1, \dots, c_m\}$. This representation follows the logic of analytic assessment rubrics, in which performance is evaluated not only holistically but also against separate criteria or dimensions of quality [22]. Each criterion c_i is assigned a normative weight w_i in the assessment rubric, reflecting that criterion's contribution to the overall result.

For each criterion (c_i), the system produces a criterion-level score $s_i(x) = \phi_i(F_i(x), R_i, E_i)$. Here, R_i is the textual description of the rubric criterion, $F_i(x)$ denotes the response features derived by the system, and E_i denotes the supporting passages on which the criterion-level score is based. This decomposition is consistent with current criterion-oriented approaches to automated assessment, which represent a student response not as a single undifferentiated text but as a set of features aligned with specific rubric items. Under this approach, rubric criteria serve not only to produce an overall score but also to train the model, monitor assessment validity, and explain the system's decisions [12].

If a criterion is scored on the ordinal scale $L_i = \{0, 1, \dots, K_i\}$, the system generates a distribution over the permissible score levels. The criterion-level score is defined as the expected score level:

$$s_i(x) = \sum_{k=0}^{K_i} k \cdot p_i(k | x), \quad (10)$$

where $p_i(k | x)$ is the normalized score indicating how well response x matches level k under criterion c_i . The resulting value is used as an intermediate criterion-level score and is not interpreted as an independently calibrated posterior probability unless a separate calibration procedure has been performed.

The overall preliminary decision is produced by normalized weighted aggregation of the criterion-level scores. The calculation incorporates both the normative criterion weights specified in the assessment rubric and the reliability of the score for each criterion:

$$S(x) = \frac{\sum_{i=1}^m w_i q_i(x) s_i(x)}{\sum_{i=1}^m w_i q_i(x)} \quad (11)$$

where w_i is the normative criterion weight, $s_i(x)$ is the criterion-level response score, and $q_i(x) \in [0, 1]$ is a reliability indicator that accounts for input quality and the sufficiency of the evidential basis.

Decision-Explanation Model

To ensure explainability, the system output is represented using the following structure:

$$J(x) = (c_i, s_i, e_i, r_i, u_i)_{i=1}^m, \quad (12)$$

where c_i is a criterion, s_i is the criterion-level score, e_i denotes supporting evidence, r_i is the textual explanation, and u_i is criterion-level uncertainty.

This representation allows an instructor to examine not only the result but also the basis for it: which criteria were applied, which features were extracted, which parts of the response served as evidence, and where the system registered uncertainty.

Model of Uncertainty and Decision Referral

Uncertainty estimation serves as a control on the reliability of the preliminary decision. A quantitative uncertainty estimate can identify cases in which a model prediction should not be regarded as sufficiently reliable for automatic use [14]. Uncertainty for each criterion is defined through the maximum posterior probability of the score level:

$$u_i(x) = 1 - \max_k P(y_i = k | x, c_i). \quad (13)$$

Aggregated uncertainty is calculated as follows:

$$U(x) = \sum_{i=1}^m w_i u_i(x). \quad (14)$$

The rule for referring a decision to an instructor is based on abstention when reliability is insufficient. If the system detects high uncertainty, poor input quality, inadequate evidence, or disagreement among criteria, the preliminary decision is not applied automatically. The student's response is instead referred to an instructor for expert review [15].

Implementation Details and Reproducibility

To make the deployed configuration independently reproducible, all components, models and thresholds of the prototype are specified here. Automatic speech recognition uses Gemini 2.5 Flash in transcription mode with word-level start/end timestamps; the pause threshold for fluency indicators is 250 ms, and a response whose transcript lacks usable word timing is flagged and its temporal indicators are reported as unavailable rather than imputed. Interpretable response features are computed by deterministic code with no learned parameters. Semantic representations use the gemini-embedding-2 model (3072 dimensions, Matryoshka-truncated to 768; early pilot turns used gemini-embedding-001), with cosine similarity computed in PostgreSQL/pgvector over an HNSW index. In the deployed assistant configuration, retrieval is top-k cosine search over course-material chunks with a lexical full-text fallback when embeddings are unavailable; these correspond to the limiting cases of the hybrid relevance (1). Course materials are chunked by a heading-aware procedure (approximately 500-character target, 800-character maximum) and each chunk is

embedded with a contextual heading-path prefix. Generated citation indices in assistant answers are machine-validated against the actually retrieved context, and hallucinated indices are dropped before the answer is shown; the citation rate reported in Table 1 is therefore an attribution-coverage indicator, not evidence of semantic groundedness, which the measure (3) formalizes and whose empirical evaluation is deferred to the validation study.

The criterion-judgment stage uses Gemini 2.5 Flash-Lite prompted critique-first: the model must first enumerate what is missing with respect to each rubric criterion and only then score. The rubric is placed in a stable prompt prefix, and the student response is wrapped in untrusted-content delimiters as prompt-injection mitigation. The output is validated against a JSON schema; the weighted aggregate (11) is computed deterministically by the system and is never taken from the model. If any per-criterion confidence falls below 0.6, the response is flagged and re-graded by Gemini 2.5 Flash. On provider failure, the request falls back to DeepSeek. The model identifier and version are logged with every score; observed low-confidence re-grading and failover frequencies are reported separately in the Results section. The IELTS-format rubric contains five criteria scored in bands 1–5, with equal normative criterion weights $w_c=0.20$. The released grade is computed as $S_{\text{final}}=\lambda S_{\text{AI}}+(1-\lambda)S_{\text{instructor}}$, with default $\lambda=0.5$, adjustable by the instructor. Only the instructor can release the result. For every indicator the system stores the raw inputs, the formula, a step-by-step derivation and the response excerpts used as evidence, forming the audit trail.

Ethics and consent. Before the examination each student completed an in-system informed-consent step (available in English, Russian and Kazakh) covering recording, automated preliminary analysis and instructor review. Records are pseudonymized; audio recordings are retained for at most 30 days and then deleted. AI outputs were preliminary throughout the pilot: final grades were assigned and released by instructors. Survey participation was voluntary and did not affect grades.

Pilot Design and Processing of Survey Data

The pilot evaluation of the AI-examiner mode included a student survey administered after participants had interacted with the system in an examination setting. The initial dataset contained 58 records; test and administrative entries were excluded, leaving an analytical sample of 52 questionnaires. Items were rated on a five-point Likert scale. Higher values on positively worded items indicated a more favorable perception of the procedure, whereas higher values on negatively worded items indicated greater difficulty or risk [23].

For the AI-assistant mode, the analysis used system interaction logs containing request-processing status, response time, source-citation status, and the results of relevant-context retrieval. A request was considered successful when an answer was produced without a technical error. The completion rate was calculated as the number of successfully processed requests divided by the total number of requests.

For each questionnaire item, the mean, its 95% confidence interval based on Student's t-distribution, and the proportion of responses in each scale category were calculated. Negatively worded items were reverse-coded before composite indices were computed, placing all items on a common interpretive direction: higher values indicate a more favorable perception of the system.

Results

Technical Results for the AI-Assistant Mode

The AI-assistant mode was evaluated using interaction logs, response-time data, request statuses, the presence of source citations, and the quality of relevant-context retrieval (Table 1).

Table 1. Technical indicators for the AI-assistant mode.

Indicator	Value	Interpretation
Number of requests	347	pilot usage volume
Successfully processed requests	336	completed responses
Request completion rate	96.8%	proportion of requests completed without a recorded technical error
Responses with source citations	89.0%	validated citation coverage of retrieved context; not a measure of semantic groundedness
Mean response time	1,247 ms	responsiveness
Median response time	980 ms	typical interaction time
P90	2,847 ms	upper-range latency

Implementation Status of the AI Examiner

The functional readiness of the AI-examiner mode was evaluated in terms of the implementation of the principal stages of criterion-based analysis: receipt of the examination task, rubric, and student response; transcription of an oral response; segmentation; feature extraction; generation of a preliminary score; identification of strengths and weaknesses; and registration of problematic passages and uncertainty signals. The implemented components and the methods used for their functional verification are presented in Table 2.

Table 2. Implemented components of the AI-examiner mode and methods of functional verification.

Component	Implemented Function	Method of Functional Verification
Input-data processing	Receipt of the examination task, assessment rubric, and student response	Verification that input data are correctly uploaded, stored, and transferred among system components
Automatic speech recognition	Conversion of the student's audio response into a text transcript	Analysis of stored transcripts and speech-recognition metadata
Response segmentation	Division of a transcript or written response into meaning-bearing units	Inspection of the system's intermediate structured output
Criterion-based analysis	Generation of criterion-level scores for individual rubric criteria	Analysis of the structured result containing a score for each criterion
Extraction of supporting evidence	Identification of response passages on which criterion-level scores are based	Inspection of stored response passages linked to the relevant criteria
Uncertainty estimation	Generation of an indicator or signal that the decision may be insufficiently reliable	Inspection of the uncertainty value and recorded reason for referring the response for additional review
Instructor verification	Review, confirmation, or revision of the system's preliminary judgment	Comparison of the AI's initial judgment with the instructor's final decision in the audit trail

Functional testing showed that the system passes a response sequentially through the main processing stages and produces a structured output containing criterion-level scores, supporting evidence, and uncertainty signals. The existence of an implemented component and its successful execution of the intended function demonstrate technical readiness, but do not establish the accuracy, validity, or pedagogical reliability of the resulting scores.

The current pilot data should therefore be interpreted as evidence of the technical pipeline's feasibility and as an assessment of user perceptions, not as conclusive evidence of automated-scoring accuracy. Full validation will require a corpus of responses independently annotated by expert instructors, comparison of the system's criterion-level and overall scores with instructor scores, measurement of inter-rater agreement, and calculation of the proportion of cases in which instructors confirm or revise the AI's preliminary judgment.

Measured Feature Distributions over Assessed Oral Responses

Beyond functional verification, Table 3 reports the measured distributions of the interpretable indicators of the feature-extraction stage over the 45 oral responses fully processed by the pipeline in the IELTS-format pilot. Word-level timing was available for 33 of 45 responses (73.3%); for the remaining 12 (26.7%) the ASR returned no usable word timestamps, so temporal indicators are reported as unavailable and the response is flagged for manual review. This provides a quantified operational consequence of missing ASR timing information: the affected temporal features cannot be computed, so the fluency block is suppressed and the response is referred for review under (13)–(14). The present pilot does not quantify how transcription errors alter criterion-level or aggregate assessment scores; score-level error propagation remains part of the planned validation study. Four articulation-rate values outside a physiologically plausible band (30-400 words/min), produced by degenerate timestamps, were excluded and are counted in the exclusion column.

Table 3. Measured distributions of feature-extraction indicators over assessed oral responses (n = 45 responses; per-indicator n varies with timing availability; Excl. = implausible values excluded under documented bounds).

Indicator	n	Excl.	Median	IQR	Mean	SD
Speech rate, words/min	31	2	78.0	66.2-102.0	87.4	42.9
Articulation rate, words/min	25	4	215.8	187.2-225.6	219.9	48.6
Phonation-time ratio	33	0	0.30	0.16-0.42	0.31	0.24
Pause count (>250 ms)	33	0	15	1-40	28.1	34.9
Mean pause duration, s	25	0	0.83	0.65-1.37	1.33	1.50
Mean length of run, words	31	2	2.8	2.0-5.0	4.9	5.4
Type-token ratio	45	0	0.79	0.65-0.93	0.77	0.17
Guiraud index	45	0	3.77	3.53-6.91	4.98	2.40
MTLD	45	0	53.6	19.0-63.0	52.2	37.7
Lexical density	45	0	0.53	0.50-0.59	0.59	0.18
Mean length of sentence, tokens	45	0	15.0	9.5-16.6	14.3	7.2
Mean length of T-unit, tokens	45	0	9.6	8.5-15.0	10.5	4.8
Prompt relevance, cosine (9)	45	0	0.57	0.52-0.75	0.63	0.14
Local coherence, adjacent cosine	25	0	0.67	0.63-0.70	0.67	0.07

Across the 215 scored turns, 13 (6.0%) triggered the low-confidence rule and were re-graded by Gemini 2.5 Flash, while 9 (4.2%) required provider failover. Separately, 12 of the 45 oral responses analyzed in Table 3 (26.7%) lacked usable word-level timing and were referred for manual review. These percentages therefore describe different denominators and different control events: model-confidence re-grading, provider failover, and ASR-timing referral, respectively.

A Worked End-to-End Example

To connect the formal model (4)-(14) with the empirical pipeline, Table 4 traces one real (pseudonymized) student response from the IELTS-format pilot through every stage: the oral answer to a cue card about technology solving a problem was transcribed with word timings, features were extracted, five criterion scores with evidence excerpts and confidence values were produced, the weighted aggregate was computed deterministically, and the instructor reviewed and blended the result. The AI preliminary score (56/100) and the instructor score (100/100) disagreed substantially; under the default blend ($\lambda = 0.5$) the released grade was 78/100, and the disagreement was logged in the audit trail.

Table 4. A real student response traced through the complete assessment pipeline, including feature extraction, criterion-based analysis, evidence selection, uncertainty estimation, preliminary decision, and instructor verification

Pipeline stage	Recorded value (real case)
Input	Oral response, IELTS-format cue card (technology that helped solve a problem); consent recorded
ASR transcript (excerpt)	"...I was preparing a presentation for my students. But the next time when I checked... I speak about I use them. I use them. I use them."
Extracted features	speech rate 68.2 words/min; TTR 0.620; Guiraud 7.26; prompt-relevance cosine 0.746
Criterion scores (bands 1-5, confidence)	fluency-coherence 2 (0.9); lexical resource 3 (0.9); grammatical accuracy 3 (0.9); task achievement 3 (0.9); structure 3 (0.9)
Evidence excerpts	repetition pattern "I use them. I use them. I use them." cited for fluency-coherence and structure; on-topic excerpt cited for task achievement
Model / escalation	initial low-confidence critique triggered escalation; final judgment by Gemini 2.5 Flash (logged)
Deterministic aggregate (11)	AI preliminary score 56/100
Instructor verification	instructor score 100/100; blended release ($\lambda = 0.5$): 78/100; AI-instructor disagreement logged

A second real case illustrates the referral mechanism: a response delivered entirely in Russian to an English-language task produced criterion scores with the explicit marker "no supporting evidence" for every criterion, an AI preliminary score of 20/100, and a mandatory-review flag; the instructor nevertheless released 100/100. Such cases are exactly what the audit trail is designed to expose for moderation.

Student Perceptions of the AI Examiner

Students' perceptions of the AI-examiner mode were examined using questionnaires completed after the system had been piloted in an examination setting (Fig. 5). After test and administrative records were excluded, the analytical sample comprised 52 responses. The questionnaire contained positively worded items addressing procedural clarity, the usefulness of criterion-level feedback, sufficiency of response time, perceived fairness, and willingness to use the system, as well as negatively worded items concerning uncertainty about the grade, the risk of speech-recognition errors, discomfort with AI involvement, and technical failures. Values are reported as mean scores on a five-point Likert scale with 95% confidence intervals. Blue bars represent positively worded items; orange bars represent negatively worded items capturing potential risks and limitations.

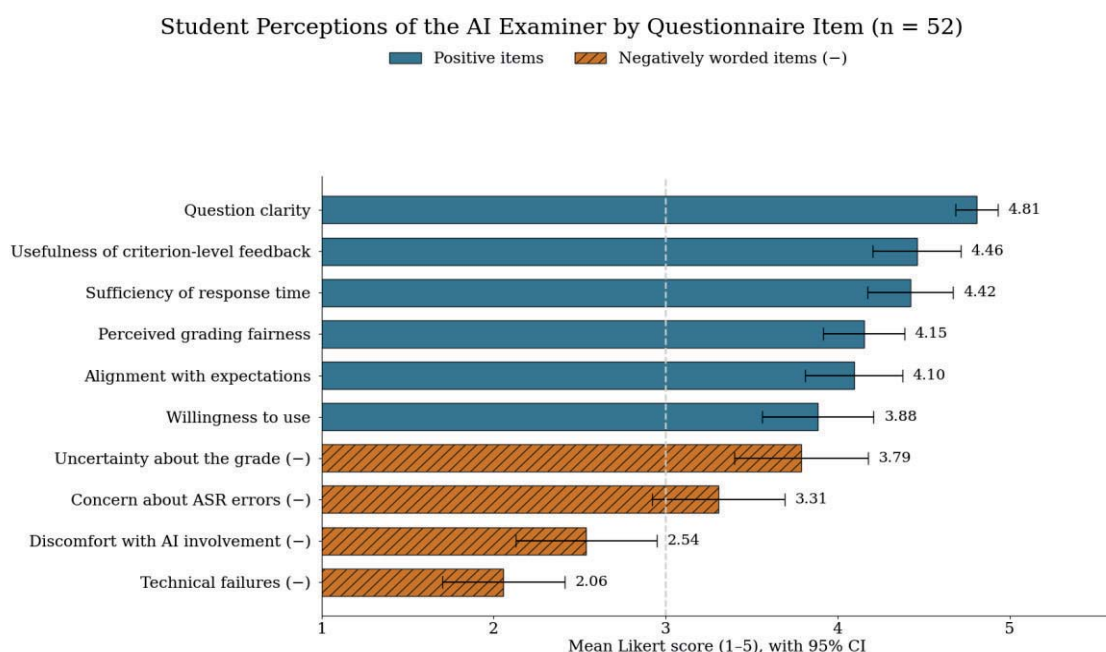


Figure 5. Mean student ratings of the AI examiner by questionnaire item (n = 52; five-point Likert scale; 95% confidence intervals)

The highest ratings were recorded for question clarity (4.81), the usefulness of criterion-level feedback (4.46), and the sufficiency of response time (4.42). These results suggest that students found the procedure for interacting with the AI examiner understandable and organizationally acceptable. The comparatively high scores for perceived grading fairness (4.15) and alignment with expectations (4.10) further suggest that the system's preliminary judgment was generally not viewed as arbitrary or opaque.

The findings nevertheless show that trust in the AI examiner is not yet unconditional. The relatively high scores on the negatively worded items 'uncertainty about the grade' (3.79) and 'concern about ASR errors' (3.31) underscore the need to retain instructor review, particularly for oral responses, where transcription quality can affect subsequent criterion-based analysis. By contrast, discomfort with AI involvement (2.54) and technical failures (2.06) were less pronounced, suggesting that the system is technically acceptable for pilot use as a preliminary analysis tool.

Taken together, the chart supports the central design principle of the proposed architecture: the AI examiner can structure a response, perform criterion-level analysis, and prepare explainable feedback, but it should not serve as an autonomous source of the final academic decision. The most defensible use case is a human-in-the-loop arrangement in which the system prepares a preliminary judgment while the instructor retains authority to review, revise, and make the final decision.

Figure 6 shows the percentage distribution of responses across the five Likert categories for each item. For positively worded items, a predominance of ratings 4 and 5 indicates favorable perceptions of the procedure. For negatively worded items, high ratings indicate that the corresponding risk or difficulty was more strongly experienced.

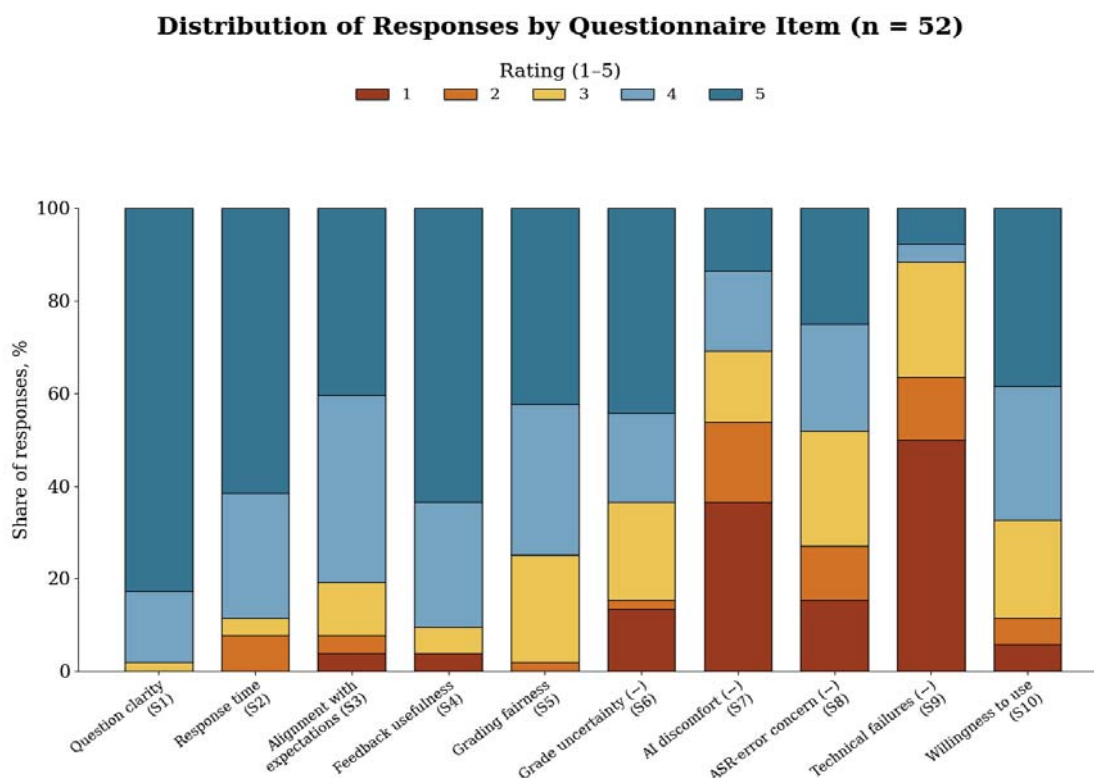


Figure 6. Distribution of student responses by questionnaire item (n = 52)

The response distributions corroborate the item means: ratings of 4 or 5 accounted for 98.1% of responses on question clarity and 90.4% on the usefulness of criterion-level feedback. Distributions were more mixed for alignment with expectations and perceived grading fairness. On the negatively worded items, the substantial share of high ratings for uncertainty about the grade and concern about ASR errors reinforces the need for instructor review and transcription-quality control in oral-response settings.

Responses to the technical-failures item, by contrast, were concentrated in the lower categories: the predominance of ratings 1 and 2 indicates that technical problems were not the principal source of negative perceptions. Further development should therefore focus less on basic operability than on strengthening confidence in the assessment logic, improving speech-recognition quality, and making instructor verification more transparent.

To provide an integrated view of students' perceptions of the AI examiner, composite indices were calculated after reverse-coding the negatively worded items. The resulting indices and their interpretation are presented in Table 5.

Table 5. Composite indices of students' perceptions of the AI examiner after reverse-coding negatively worded items.

Index	Constituent Items	Mean	Interpretation
Procedural transparency and acceptability	clarity, time, usefulness of feedback, willingness to use	4.39	the procedure is clear and generally accepted by students
Perceived fairness and alignment with expectations	alignment with expectations, fairness, reverse-coded uncertainty	3.49	confidence in the assessment judgment is emerging but remains limited

Index	Constituent Items	Mean	Interpretation
Trust in and acceptance of the AI examiner	reverse-coded discomfort, ASR concerns, and technical failures	3.37	a priority area for further improvement
Overall perception index	all items, with negatively worded items reverse-coded	3.81	the system is perceived favorably but still requires instructor oversight

The composite indices show that procedural transparency and acceptability were the strongest outcomes of the pilot (4.39). Lower scores for perceived fairness and alignment with expectations (3.49), and for trust in and acceptance of the AI examiner (3.37), indicate that the system may be used as a preliminary analysis tool but should not be treated as an autonomous mechanism for final assessment. The overall perception index (3.81) reflects a generally favorable attitude toward the system, provided that instructor oversight is retained.

Preliminary Agreement Between AI and Instructor Scores

For the 23 IELTS-format pilot sessions for which both an AI preliminary score and an instructor-released score were available, Table 6 provides a descriptive paired comparison on the common 0–100 scale. The AI scores had a mean of 33.2 (SD 26.3), whereas instructor scores had a mean of 96.5 (SD 11.5). The association was near zero (Pearson $r = 0.09$; Spearman $\rho = -0.05$), with a mean absolute difference of 63.3 points.

The instructor-score distribution exhibits a severe ceiling effect and restriction of range. Because these grades were not assigned using the same band-anchored criterion rubric used by the AI, the source of this compression cannot be determined from the pilot data. Consequently, the instructor-released grades cannot be treated as a criterion measure of AI scoring accuracy.

The comparison is therefore reported as a diagnostic result rather than as a validation result. It demonstrates why the subsequent agreement study must use independent trained raters, a common rubric, and band-anchored descriptors. The largest observed disagreements are retained for error analysis but are not interpreted as evidence that either the AI or the instructor produced the correct assessment.

Table 6. AI preliminary vs. instructor-released scores, IELTS-format pilot (0-100 scale).

Statistic	Value
Paired sessions, n	23
AI preliminary score, mean +/- SD	33.2 +/- 26.3
Instructor score, mean +/- SD	96.5 +/- 11.5
Pearson r	0.09
Spearman rho	-0.05
Mean absolute difference	63.3 points

Discussion

The central finding of this study is that the AI assistant and AI examiner can be implemented as two modes of a single intelligent system for decision support and justification. In technical terms, such a system requires more than a language model: it also needs a formalized criterion layer, an evidence layer, an uncertainty layer, logging, and a human-in-the-loop verification process.

The proposed architecture reduces the risks associated with ungoverned use of a generative model. The language model does not make the final decision; its role is restricted to text interpretation, feature extraction, generation of a structured explanation, and preparation of a preliminary judgment [11]. In consequential cases, the final decision remains with the instructor. This constraint on AI autonomy is consistent with a risk-based approach to artificial intelligence

in education [5]. Systems used to assess learning outcomes may be classified as high risk and are consequently expected to provide human oversight, transparency, interpretability, logging, and a means for a human reviewer to reject or override the system's output.

A limitation of the present stage of the study is the absence of a response corpus independently annotated by expert instructors for each rubric criterion. The pilot findings should therefore not be interpreted as evidence of automated-scoring accuracy, which requires validation against independently assigned expert ratings [24]. They establish technical feasibility, procedural clarity, and favorable user perceptions, but cannot substitute for subsequent expert validation. Future validation should compare the AI's criterion-level and overall scores with instructor ratings, analyze cases of substantial disagreement and instructor revision [25], and explicitly test whether and to what extent automatic-transcription errors propagate into criterion-level and overall assessment results [26].

The current pilot provides three forms of intermediate evidence relevant to subsequent validation: measured distributions of interpretable response features (Table 3), an end-to-end traceability example (Table 4), and a preliminary paired-score comparison (Table 6). None of these results is interpreted as evidence of scoring validity.

In the next study, a benchmark corpus of 30–50 responses will be sampled independently from eligible responses to cover the observed range of performance. Each response will be rated independently by two trained raters using band-anchored descriptors for every rubric criterion. Human–human agreement will be reported as a reliability benchmark using quadratic-weighted kappa for criterion-level ratings and intraclass correlation for the aggregate score. After the independent ratings have been completed, disagreements will be adjudicated to establish the expert reference.

AI–expert agreement will then be evaluated using quadratic-weighted kappa for criterion-level scores and ICC, MAE, and RMSE for the aggregate score. On the same benchmark corpus, an unconstrained general-purpose LLM prompt and the proposed criterion-guided pipeline will be compared against the same expert reference. Cases flagged by the deployed system will be analyzed separately as a challenge set and will not define the benchmark sample. Instructor verification time, override frequency, override direction, and stated reasons will also be logged.

From a technical perspective, the separation of probabilistic and deterministic layers is particularly important. The probabilistic layer comprises the language model, ASR, and semantic retrieval. The deterministic layer comprises the rubric, criterion weights, aggregation rules, uncertainty thresholds, referral rules, and audit trail. This separation improves reproducibility, auditability, and governance of the resulting decisions.

Conclusion

This study presented and pilot-evaluated a criterion-guided human-in-the-loop architecture that combines RAG-based learning support with preliminary analysis of written and oral student responses. The architecture separates probabilistic AI components from deterministic assessment and governance rules and produces a traceable record linking response features, supporting evidence, criterion-level scores, uncertainty handling, aggregation, and instructor verification.

The pilot establishes technical feasibility, procedural acceptability, and the operational behavior of the proposed control mechanisms. Among 45 analyzed oral responses, 12 (26.7%) lacked usable word-level timing and were referred for review rather than evaluated using unsupported temporal indicators. Across 215 scored turns, 13 (6.0%) triggered low-confidence re-grading and 9 (4.2%) required provider failover. A worked example demonstrated end-to-end auditability, while the preliminary comparison of 23 paired AI and instructor-released scores showed that the available instructor grades were unsuitable as a criterion reference because of severe ceiling restriction.

These results do not establish the accuracy, reliability, or validity of automated assessment. The next stage will therefore use an independently rated, band-anchored benchmark corpus to measure human–human reliability, AI–expert agreement, aggregate prediction error, and

comparative performance against an unconstrained LLM baseline. Until such validation is completed, the AI examiner should be regarded as a decision-support and preliminary-analysis system in which final academic judgment remains with the instructor.

References

- [1]. Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education*, 16, 39. <https://doi.org/10.1186/s41239-019-0171-0>
- [2]. González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, 11(12), 5467. <https://doi.org/10.3390/app11125467>
- [3]. Manganello, F., Nico, A., & Boccuzzi, G. (2025). Theoretical foundations for governing AI-based learning outcome assessment in high-risk educational contexts. *Information*, 16(9), 814. <https://doi.org/10.3390/info16090814>
- [4]. Fajardo-Ramos, D. C., Chiappe, A., & Mella-Norambuena, J. (2025). Human-in-the-loop assessment with AI: Implications for teacher education in Ibero-American universities. *Frontiers in Education*, 10, 1710992. <https://doi.org/10.3389/feduc.2025.1710992>
- [5]. Khan, W., Topham, L., Jones, N., Atherton, P., Al-Shabandar, R., Kolivand, H., Khan, I., Alatrany, A., & Hussain, A. (2026). Auto-assessment of assessment: A human-in-the-loop AI framework addressing policy gaps in academic assessment. *PLOS ONE*, 21(4), e0346815. <https://doi.org/10.1371/journal.pone.0346815>
- [6]. Mousavinasab, E., Zarifsanaiey, N., Niakan Kalhori, S. R., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2021). Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, 29(1), 142–163. <https://doi.org/10.1080/10494820.2018.1558257>
- [7]. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [8]. 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. *Advances in Neural Information Processing Systems*, 33, 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf
- [9]. Stolfo, A. (2024). Groundedness in retrieval-augmented long-form generation: An empirical study. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 1537–1552). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-naacl.100>
- [10]. Emirtekin, E. (2025). Large language model-powered automated assessment: A systematic review. *Applied Sciences*, 15(10), 5683. <https://doi.org/10.3390/app15105683>
- [11]. Xie, W., Niu, J., Xue, C. J., & Guan, N. (2026). Grade like a human: Rethinking automated assessment with large language models. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems (RACS '25)* (Article 21, 8 pages). Association for Computing Machinery. <https://doi.org/10.1145/3769002.3769962>
- [12]. Kumar, V. S., & Boulanger, D. (2021). Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education*, 31(3), 538–584. <https://doi.org/10.1007/s40593-020-00211-5>

- [13]. Ercikan, K., & McCaffrey, D. F. (2022). Optimizing implementation of artificial-intelligence-based automated scoring: An evidence centered design approach for designing assessments for AI-based scoring. *Journal of Educational Measurement*, 59(3), 272–287. <https://doi.org/10.1111/jedm.12332>
- [14]. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarencov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, 243–297. <https://doi.org/10.1016/j.inffus.2021.05.008>
- [15]. Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)* (pp. 4885–4894). Curran Associates Inc. <https://dl.acm.org/doi/10.5555/3295222.3295241>
- [16]. Rožanec, J. M., Novalija, I., Zajec, P., Kenda, K., Tavakoli Ghinani, H., Suh, S., Bian, S., Veliou, E., Papamartzivanos, D., Giannetsos, T., Menesidou, S. A., Alonso, R., Cauli, N., Meloni, A., Reforgiato Recupero, D., Kyriazis, D., Sofianidis, G., Theodoropoulos, S., Fortuna, B., Mladenčić, D., & Soldatos, J. (2023). Human-centric artificial intelligence architecture for Industry 5.0 applications. *International Journal of Production Research*, 61(20), 6847–6872. <https://doi.org/10.1080/00207543.2022.2138611>
- [17]. Komarov, N., Mukhanov, S. B., Bazarbekov, I. M., Zhakypbekov, S. Zh., & Sibanbayeva, S. Y. (2025). Methods of processing and analyzing big data in machine learning tasks: Approaches and prospects. *Herald of the Kazakh-British Technical University*, 1, 25–35. <https://doi.org/10.55452/1998-6688-2025-22-1-25-35>
- [18]. Kenzhebulatova, D., Grigoryeva, I., Yermakova, V., Bublikova, O., Tumanova, A., & Mukhamadiyev, K. (2025). The professional linguistic worldview of IT specialists as a component of linguistic culture. *Forum for Linguistic Studies*, 7(1), 98–111. <https://doi.org/10.30564/fls.v7i1.7742>
- [19]. Mambetov, S., Joldasbayev, S., Bykov, A., Koishybay, S., & Dossanbek, K. (2026). Methods for automatic emotion recognition in hacker forum texts. In K. Kolesnikova & M. Ipalakova (Eds.), *Proceedings of the Workshop on Cybersecurity, Infocommunication Systems and Networks 2025* (Vol. 4180, Paper 13). CEUR Workshop Proceedings. <https://ceur-ws.org/Vol-4180/Paper13.pdf>
- [20]. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
- [21]. Cui, B., Li, Y., Zhang, Y., & Zhang, Z. (2017). Text coherence analysis based on deep neural network. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2027–2030). Association for Computing Machinery. <https://doi.org/10.1145/3132847.3133047>
- [22]. Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education*, 42(3), 347–360. <https://doi.org/10.1080/02602938.2015.1111294>
- [23]. Jebb, A. T., Ng, V., & Tay, L. (2021). A review of key Likert scale development advances: 1995–2019. *Frontiers in Psychology*, 12, 637547. <https://doi.org/10.3389/fpsyg.2021.637547>
- [24]. Pecuchova, J., Benko, L., & Drlik, M. (2025). Automated grading of open-ended questions in higher education using GenAI models. *International Journal of Artificial Intelligence in Education*, 35, 3813–3846. <https://doi.org/10.1007/s40593-025-00517-2>

[25]. Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, 51(1), 201–224. <https://doi.org/10.1002/berj.4069>

[26]. Hickman, L., Langer, M., Saef, R. M., & Tay, L. (2025). Automated speech recognition bias in personnel selection: The case of automatically scored job interviews. *Journal of Applied Psychology*, 110(6), 846–858. <https://doi.org/10.1037/apl0001247>