

A Governed Workflow for LLM-Based Formative Assessment with Score Stabilization

Xiaorui Zhang^a, Yang Liu^b, Jinnan Kou^c and Hui Liu^{c,*}

^aInternational School of Information Science and Engineering, Dalian University of Technology, Dalian, China

^bSchool of Marxism, Shenyang Jianzhu University, Shenyang, China

^cSchool of Foreign Languages, Dalian University of Technology, Dalian, China

ARTICLE INFO

Keywords:

formative assessment
large language model
structured output governance
score stabilization
learning analytics

ABSTRACT

Open-ended formative assessment requires records that teachers can interpret and trust over time, yet large language model outputs remain variable across repeated calls. This study develops a governed workflow for LLM-based formative assessment of student core competencies. The method combines four competence dimensions, a JSON-only response contract, backend validation and normalization, threshold-based repeated inference with closest-pair aggregation, deterministic post-processing, and teacher review before publication. It is implemented in a classroom platform that stores traceable assessment records and supports follow-up reporting. Technical checks showed end-to-end execution and preservation of structured result objects with ensemble metadata. In a one-term pilot, 147 experimental and 145 control records met the strict confidence-analysis criteria. The confidence distribution differed significantly between groups, and within the high-confidence A-level subset the experimental group showed a larger later-stage reading gain than the control group (10.52 vs. 3.51, $p = 0.0029$), whereas writing differences were not stable across filtering choices. These results do not establish causal effects, but they indicate that governed processing can turn variable LLM outputs into reviewable records for classroom use.

1. Introduction

Competence-oriented assessment shifts evaluation beyond task accuracy toward broader capacities such as engagement, self-regulation, transfer, and moral reasoning. Frameworks such as the Core Competencies for Chinese Students' Development and the OECD Learning Compass 2030 both argue that educational judgement should rest on interpretable evidence of multidimensional development rather than on a single score [1, 2]. In routine classrooms, however, open-ended work is difficult to assess at scale [3]. Teacher judgement is indispensable, but sustaining it as a continuous cycle of diagnosis, feedback, goal setting, and follow-up is time-intensive and uneven across tasks.

Large language models (LLMs) offer one possible technical basis for supporting this work. Recent studies show that rubric-guided prompting and structured criteria can improve agreement with human judgement in writing evaluation and criterion-based grading [4, 5, 6, 7]. Yet classroom deployment demands more than benchmark scoring: results must be stable enough to store, review, visualize, compare over time, and release to students without exposing raw model variability. A model that scores well in isolation is therefore not enough [8]. The practical question is whether variable

generations can be turned into assessment records that remain interpretable, reviewable, and reusable in ordinary teaching.

Against this background, the present study proposes a governed workflow for LLM-based formative assessment intended to inform classroom judgement about student core competencies. The workflow begins with a student's submitted assignment response. It organizes evidence along four competence dimensions—learning attitude, learning ability, learning strategies, and values with moral reasoning—and routes model outputs through contract enforcement, normalization, stabilization, deterministic post-processing, and teacher review before publication. The result is a classroom-oriented workflow in which LLMs provide preliminary judgement while pedagogical control remains explicit and traceable.

Three contributions define the proposed method:

- It operationalizes four student core-competence dimensions into an evidence-oriented assessment setup that is compatible with LLM-supported formative evaluation rather than with one-shot automated scoring alone.
- It proposes a governed evaluation pipeline that combines rubric-aligned prompting, a JSON-only output contract, validation and normalization, repeated inference, aggregation, deterministic post-processing, and a teacher-review checkpoint.
- It reports technical verification and pilot evidence showing how the governed workflow can be deployed as a usable classroom method while keeping claims about instructional effect appropriately cautious.

DOI: <https://doi.org/10.70891/JAIR.2026.030059>

ISSN of JAIR: 3078-5529

License: CC-BY 4.0, see <https://creativecommons.org/licenses/by/4.0/>

4.0/

*Corresponding author

xiaorui537537@gmail.com (X. Zhang); sjmks@sjzu.edu.cn (Y. Liu);

koujinnan@dlut.edu.cn (J. Kou); liuhui1126@dlut.edu.cn (H. Liu)

Accordingly, the empirical scope of the paper is workflow feasibility and aggregate classroom interpretability in assignment-based use. The pilot does not directly establish the psychometric validity of the four dimension scores as standalone measures.

2. Related Work

Recent work on LLMs in educational assessment has developed along several partially connected lines. The first centers on open-ended writing evaluation and criterion-based grading. Studies on automated essay scoring consistently report that explicit rubrics, exemplars, and multidimensional criteria can improve alignment between LLM outputs and human judgements across content, organization, and argumentation [4, 5, 6, 7]. This literature establishes that free-form student responses can be scored with more structure than earlier surface-feature approaches, but it still focuses mainly on writing quality itself rather than on broader learner characteristics such as strategy use, engagement, or values-oriented reasoning.

The second line concerns robustness in short-answer and process-oriented evaluation. Results here are more mixed. LLMs can support grading of short responses and explanations, yet their performance is less dependable when responses are partially correct, underspecified, or locally atypical [9, 10]. Such findings have encouraged workflows in which automated scoring is accompanied by explanation checks, re-prompting, or teacher sampling rather than being treated as a one-pass replacement for classroom judgement. The implication is that usefulness depends not only on benchmark agreement, but also on how uncertainty is handled in the deployment workflow.

The third line focuses on fairness, psychometric reliability, and bias auditing. Recent studies show that acceptable mean agreement is not sufficient for dependable educational use, especially when outputs contribute to feedback, reporting, or higher-stakes decisions [11, 12]. Bias-audit research further argues that evaluation should examine whether model outputs remain balanced when subgroup membership is irrelevant to task performance [13]. Taken together, these studies shift attention from “can the model score?” to “under what governance conditions can the score be trusted?” This shift is particularly relevant when records are meant to be stored, reviewed, and revisited over time.

Another strand emphasizes structured outputs, assessment redesign, and teacher assessment literacy in the era of generative AI. Rubric embedding, constrained output formats, and explicit reasoning stages can make LLM generations easier to validate and reuse downstream [4, 5]. However, much of this literature remains focused on redesigning tasks or policies at a general level, while giving less attention to how structured outputs become reusable assessment records inside an actual classroom system.

The final line concerns formative assessment and educational support in routine teaching. Classroom formative-assessment research emphasizes that effective feedback is interactive, teacher-mediated, and tied to co-regulated or

self-regulated learning rather than to one-off score release [14, 15]. In parallel, LLM-based educational agents have shown potential for diagnosis, reflection, goal setting, and ongoing support [16, 17, 18]. What remains less developed is the bridge between these strands: a governed classroom workflow that converts variable LLM generations into stable, traceable, and teacher-reviewable assessment records. The present study addresses that gap by focusing on the runtime path from model generation to publication-ready classroom artifacts.

3. Method

3.1. Competence-Oriented Assessment Setup

The system is designed as a classroom-oriented formative-assessment platform rather than as a standalone scoring script. Figure 1 summarizes the core governed assessment method. Student, teacher, and administrator portals share a common business and orchestration layer that hosts assignment handling, AI assessment, behavior insight, speaking practice, and goal tracking. For the method proposed here, the key design choice is that governance sits beside model invocation rather than after it. Structured output governance, validation and normalization, score stabilization, the teacher-review checkpoint, and metadata recording all lie on the runtime path.

Assignment responses provide the primary evidence for competence assessment. The model is asked to interpret the submission through four operational dimensions summarized in Table 1. These dimensions are not introduced as a new psychometric instrument. Instead, they function as a pragmatic decomposition that translates broad educational constructs into observable evidence categories suitable for rubric-aligned prompting. The decomposition draws on the Chinese core-competence framework, the OECD Learning Compass, and related work on engagement, self-regulated learning, transfer, and moral argumentation [1, 2, 15, 19, 20, 21, 22].

3.2. Structured Output Contract and Normalization

Each evaluation request is framed as a rubric-aligned prompting task. The input stage segments the submission into evaluable text units and pairs them with dimension-relevant criteria, scoring anchors, and output instructions. Rather than allowing an unconstrained narrative judgement, the system requires a JSON-only response containing dimension-level judgements, supporting evidence, actionable suggestions, and an overall object. In the verified runtime, published records preserve stable overall scores, dimension averages, holistic feedback, and ensemble metadata in machine-readable form.

This contract serves two purposes. First, it reduces ambiguity between what the model is expected to observe and what the downstream platform can actually store and display. Second, it makes governance possible. Backend validation checks whether the response matches the expected schema, whether numeric fields lie within the allowed range, and

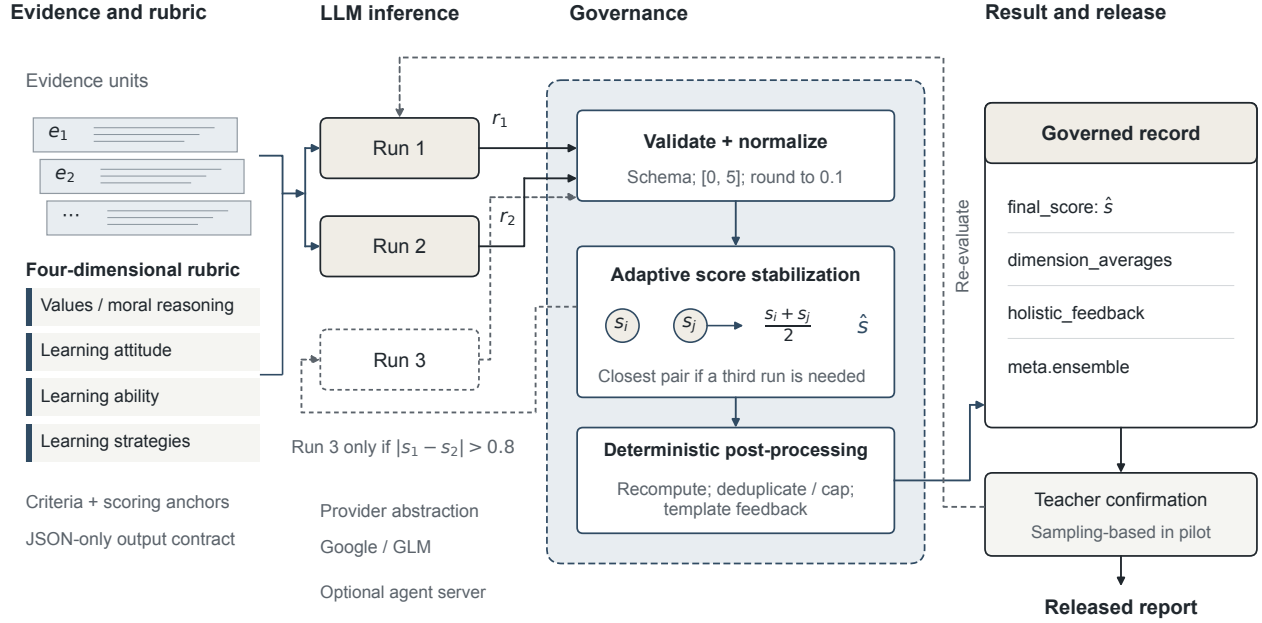


Figure 1: Core governed assessment method. Submission evidence and the four-dimensional rubric enter two default model runs. Schema validation and score normalization precede adaptive stabilization; a third run is requested only when $|s_1 - s_2| > 0.8$. Deterministic post-processing produces a governed record for teacher confirmation before release. Dashed paths indicate conditional evaluation or re-evaluation. Evidence strips are schematic.

Table 1

Operational decomposition of the four competence dimensions.

Dimension	Evidence focus	Interpretive emphasis
Values with moral reasoning	Positions, moral concerns, and engagement with counter-positions	Breadth of value considerations and structural maturity of ethical argumentation
Learning attitude	Emotional engagement, persistence under difficulty, and task focus	Meaningful engagement and sustained effort during learning
Learning ability	Higher-order cognition, metacognitive monitoring, and transfer	Analysis, evaluation, creation, reflective regulation, and application in new contexts
Learning strategies	Strategy selection, combination, and adaptation	Diversified and reflective strategy use instead of surface-level completion

whether required sections are present before the result can enter later stages. Outputs that fail these checks are not released directly to students.

Normalization is applied before any aggregation step. Scores are clamped to the system range of 0–5 and rounded to one decimal place so that later comparisons operate on standardized values rather than on raw free-form generations. Evidence items and suggestions are also normalized into stable lists. In this way, structured prompting is treated not merely as a prompt-writing technique, but as part of a machine-readable assessment contract.

Table 2 summarizes the publication-oriented contract used in the current implementation. Each field has both a semantic role and a validation rule. Overall scores and dimension averages are not accepted as opaque model outputs; they

are checked and, when necessary, recomputed deterministically. Evidence excerpts and suggestions are normalized into capped lists so that reports remain readable and comparable across tasks. Ensemble metadata is preserved for traceability rather than discarded after aggregation.

The stabilized result object retains a small set of release-critical fields rather than the raw completion. In the current schema, a published record preserves the final score, dimension averages, a short holistic-feedback string, and ensemble metadata including the aggregation method, confidence score, and pairwise disagreement. These fields support reporting, audit, and later comparison without exposing raw model completions.

Table 2

Publication-oriented structured output contract.

Output block	Representative fields	Governance role
Overall assessment	final_score, dimension_averages, holistic_feedback	Range validation, deterministic recomputation, and direct use in reports and profile views
Dimension-level result	score, evidence excerpts, rationale, suggestions	Schema validation, normalization of list lengths, and teacher-review readability
Traceability metadata	model, prompt_version, method, pairDiff, confidence, runs	Supports regeneration, auditability, and explanation of stabilized results

3.3. Score Stabilization and Deterministic Post-Processing

Output variance remains a central problem even under a constrained contract. Repeated calls under the same rubric can still produce slightly different scores or evidence selections. In classroom settings, such instability matters because results are reused in reports, profile views, and longitudinal comparisons. The method therefore treats the LLM as a variable upstream component and adds a stabilization stage before publication.

Figure 2 summarizes the procedure. Two model runs are executed by default. After normalization, their pairwise difference is checked against a threshold $\tau = 0.8$. If the disagreement is within the threshold, the final score is the average of the two runs. Otherwise, the system triggers a third run, selects the closest pair among the three normalized outputs, and averages that pair. This rule is formalized by:

$$\hat{s} = \begin{cases} \frac{s_1 + s_2}{2}, & \text{if } |s_1 - s_2| \leq \tau, \\ \frac{s_i + s_j}{2}, & \text{if } (i, j) = \arg \min_{a < b} |s_a - s_b|. \end{cases} \quad (1)$$

Aggregation is followed by deterministic post-processing. The backend recomputes derived fields such as dimension averages and the final score, merges and deduplicates evidence items, caps repetitive suggestions, and produces holistic feedback through a deterministic template rather than by concatenating raw model text. This step keeps the published result internally consistent even when upstream generations differ at the wording level. To preserve traceability, the system stores ensemble metadata such as the chosen method, pairwise difference, confidence, and run history in `meta.ensemble`.

3.4. Governed Publication Workflow

The workflow proceeds through student submission, input organization, AI-supported preliminary evaluation, controlled processing, teacher review, and result publication. Raw model output is never exposed directly to student-facing views. Instead, a teacher inspects the stabilized result, triggers regeneration or re-evaluation when necessary, and publishes only the confirmed record. In the pilot deployment, this review was usually sampling-based rather than exhaustive, with attention concentrated on borderline or

ambiguous cases that benefited from organized evidence excerpts and dimension summaries. This design separates model generation from pedagogical release.

The same principle extends to downstream modules. Assessment reports feed later visualization and follow-up functions, while behavior insight is generated as an auxiliary, non-scoring interpretation layer based on bounded evidence snapshots rather than as a second grading channel. The aim is to support formative assessment in routine teaching: LLMs provide structured preliminary judgement, but authority over release, interpretation, and instructional response remains explicit and reviewable.

Figure 3 focuses on the teacher-controlled release checkpoint and the re-evaluation loop. The result that reaches the teacher is already a governed data object rather than a raw completion. This distinction matters because the system is expected to support regeneration history, student-facing reporting, and later trend analysis. Teachers therefore interact with an object whose structure has already passed validation and stabilization.

Although the present paper focuses on assignment-based assessment, the surrounding system also constrains related modules in the same spirit. Behavior insight is generated through a two-stage process in which backend code first produces a deterministic summary of observable activity and the model then generates interpretation only on the basis of that summary. Numeric behavior scoring is explicitly prohibited. This separation keeps the central assessment channel and the auxiliary interpretive channel from being conflated. These auxiliary paths provide platform context only and are not the primary empirical validation target of the present paper.

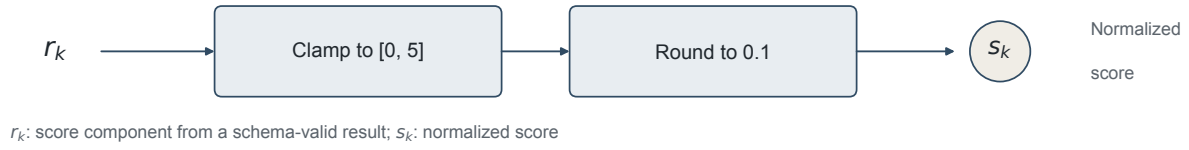
In practical terms, release readiness is not decided by the model alone. The workflow in Figure 3 places teacher confirmation after schema validation, normalization, stabilization, and deterministic recomputation, so traceability is preserved from raw generation to published report.

4. Experimental Results

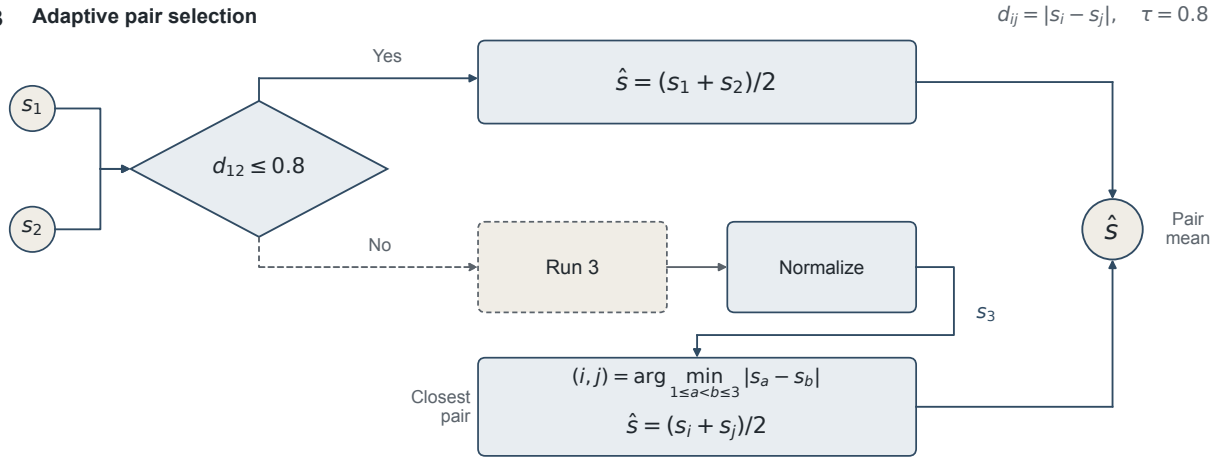
4.1. Setting and Verification Scope

The system was piloted for one academic term in a general-education course taught by Prof. Hui Liu at Dalian University of Technology. During the pilot, the experimental group used the platform for assignment submission,

A Normalize each model score before comparison



B Adaptive pair selection



C Deterministic post-processing



Figure 2: Score stabilization and deterministic post-processing. (A) Each schema-valid raw score component r_k is clamped to $[0, 5]$ and rounded to 0.1, yielding s_k . (B) The two default normalized runs are averaged when $|s_1 - s_2| \leq 0.8$. Otherwise, a third run is requested and normalized, and the closest pair among the three runs is averaged. (C) Scores are recomputed, evidence is deduplicated, suggestions are deduplicated and capped, and template feedback is generated. Ensemble metadata is retained for traceability.

AI-supported preliminary evaluation, teacher review, report publication, and follow-up access, while the control group did not receive the same system-supported workflow. The assignments were primarily open-ended writing and reflective tasks, which matched the evidence assumptions of the governed assessment method.

The evidence reported here has two layers. The first is technical verification of the implemented workflow in the local deployment environment. Front-end type checking and production build completed successfully, the backend package build passed, and live grading history records preserved structured result objects together with ensemble metadata. The second layer is pilot evidence from classroom use of the assignment-based assessment channel. Teachers mainly relied on sampling-based review, using regeneration or re-evaluation for borderline or ambiguous cases before release, while students revisited published reports through profile views, trend displays, and structured feedback. The study is reported at the level of course deployment and aggregate comparison rather than at the level of individual student

cases. No personally identifying student information is required to understand the methodological claim advanced here.

The student-row set contained 149 experimental and 150 control records. Stagewise score comparisons used pairwise complete rows for the relevant assessments, and the stricter confidence-analysis subset required valid mid-term reading, mid-term writing, final reading, and final writing records, leaving 147 experimental and 145 control observations. Between-group stage comparisons and gain comparisons are reported with Welch’s t -tests and Hedges’ g , within-group stage changes with paired t -tests and 95% confidence intervals, and confidence distributions with chi-square and Fisher tests.

Table 3 condenses the verified evidence most relevant to the methodological claim of the paper. These checks do not prove educational effectiveness by themselves, but they do show that the assessment pipeline can be executed, stored, retrieved, and reviewed as a real system path rather than as a paper-only proposal.

These results should still be read cautiously. The reported verification does not constitute a production-grade

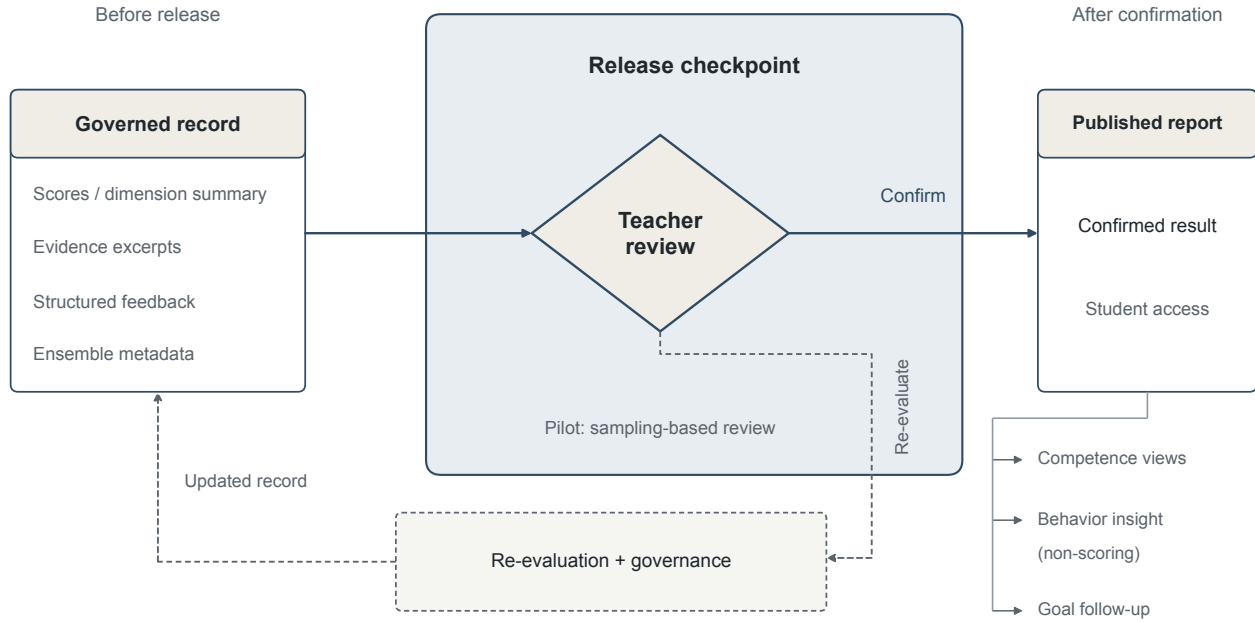


Figure 3: Teacher-controlled release of governed assessment records. The record entering teacher review has already passed validation, normalization, stabilization, and deterministic post-processing. Confirmation permits publication and student access. The dashed path indicates re-evaluation followed by governance and return of an updated record for review. Pilot review was sampling-based. Released reports support competence views and goal-related follow-up; behavior insight remains an auxiliary, non-scoring interpretation module.

Table 3
Method-relevant technical verification evidence.

Aspect	Verified evidence	Status
Local runtime readiness	Front-end type check and production build passed; backend package build passed; local frontend, backend, and database services started successfully	Pass
Assessment path	Live submission returned valid records; grading history stored structured results with overall fields and meta.ensemble; regeneration history was preserved	Pass
Behavior insight path	Stage-1 summary retrieval and Stage-2 structured insight generation were both verified in the local runtime	Pass
Speaking-practice path	Session persistence and WebSocket completion were verified, though some fresh transcript fields remained inconsistent in retest	Partial

security audit or a cross-platform deployment study, and the pilot is exploratory rather than causal. The purpose of this section is therefore to test practical feasibility and interpretability, not to claim fully generalizable instructional effects.

4.2. Pilot Findings

Reading was tracked across baseline, mid-term, and final stages, whereas writing was compared only between mid-term and final stages. In the full student sample, baseline reading did not differ significantly between groups (experimental: 59.04, control: 61.99, $p = 0.182$). Mid-term reading, however, was much higher in the control group (81.44 vs. 63.68, $p < 0.001$, $g = -1.01$), while final reading converged again (71.52 vs. 73.24, $p = 0.259$). Mid-term and

final writing scores likewise showed no significant between-group differences ($p = 0.200$ and $p = 0.470$, respectively). Figure 4 presents these reading trajectories in a dual-panel form: the full sample in Panel A and the A-level high-confidence subset in Panel B.

Within-group comparisons sharpen the same pattern. In the experimental group, reading rose from baseline to mid-term by 4.63 points ($p = 0.0046$) and from mid-term to final by another 7.67 points ($p < 0.001$). In the control group, the early reading gain was much larger at 19.49 points ($p < 0.001$), but it was followed by a significant decline of 8.28 points from mid-term to final ($p < 0.001$). Writing change was small in the experimental group and positive but modest in the control group, which already suggests that the clearest contrast lies in reading rather than writing.

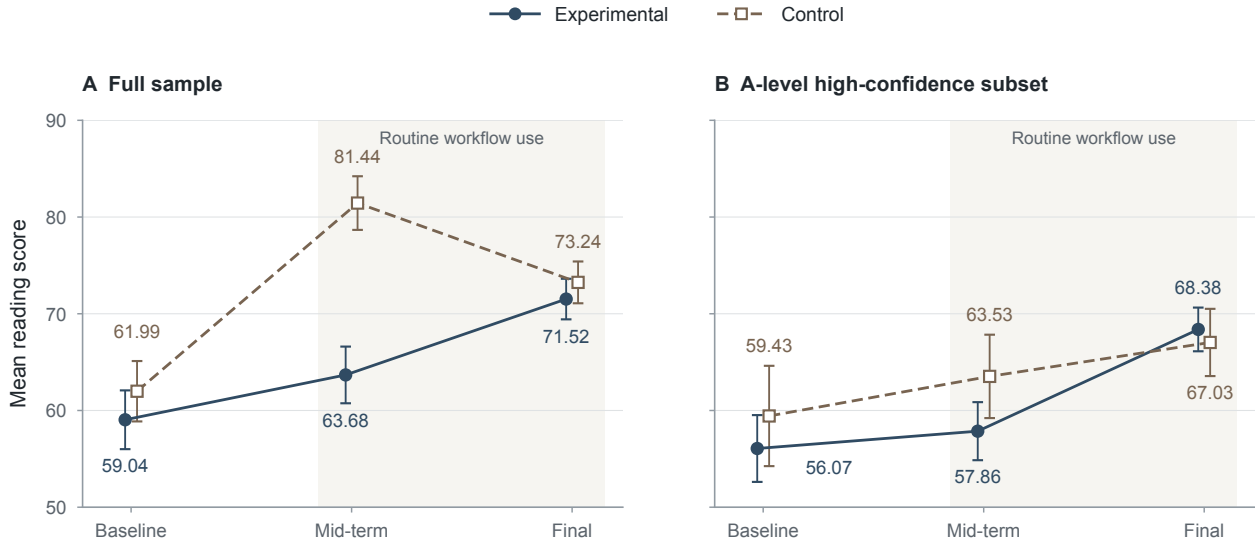


Figure 4: Reading trajectories in the full sample (A) and the A-level high-confidence subset (B). Points and labels show stagewise means; whiskers show pointwise 95% Student *t* confidence intervals. Sample sizes at baseline/mid-term/final are: full sample, experimental 145/149/148 and control 145/146/150; A-level subset, experimental 105/107/107 and control 49/50/50. Groups are slightly offset horizontally for clarity. Shading marks routine workflow use. Zero scores are treated as missing. A-level membership follows the recalibrated consistency rule among complete mid-term/final reading and writing records. Lines connect group means with varying stagewise sample sizes; gain comparisons in Table 6 use pairwise complete records.

This stagewise pattern should nevertheless be interpreted cautiously. The much higher mid-term reading mean in the control group cannot be treated as simple counter-evidence to the proposed workflow, because the pilot was conducted under ordinary classroom conditions rather than fully standardized administration. The intervention also centered on the mid-term period rather than preceding every measured score change. Under such conditions, unusually large intermediate gains may reflect instructional rhythm, local testing conditions, and short-term fluctuation rather than a stable estimate of later-stage learning development. The full-sample comparison is therefore informative, but not sufficient on its own.

To reduce the influence of potentially unstable observations, the pilot analysis introduced a confidence-aware filtering rule. Its role is methodological rather than punitive: it was designed as a group-level interpretive safeguard for aggregate comparison, not as a judgement about individual students. Calibration used student records only, excluding summary rows and treating zero exam scores as missing. At each stage (mid-term and final), writing scores were stratified into $[0, 60)$, $[60, 70)$, $[70, 80)$, $[80, 90)$ and $[90, 100]$. Within each stratum, the experimental group's valid reading scores were used to calculate the empirical 15th and 85th percentiles by linear interpolation. Table 4 reports the mid-term calibration values; final-stage thresholds were calculated separately using final writing and reading scores. The same stage-specific thresholds were applied to both groups. A stage was flagged when its reading score exceeded the corresponding 85th percentile; the 15th percentile was retained for descriptive context and was not an exclusion criterion.

Among records with valid mid-term and final reading and writing scores, cases were classified as A if neither stage was flagged, B if one stage was flagged, and C if both were flagged. These heuristic consistency categories are distinct from the 95% confidence intervals for mean scores in Figure 4. High-confidence A-level cases accounted for 72.8% of the experimental group but only 34.5% of the control group, while partial inconsistency and low-confidence cases were much more concentrated in the control group, as shown in Table 5. The three-level distribution differed significantly between groups ($\chi^2(2) = 43.39$, $p < 0.001$). When collapsed to A-level versus non-A-level cases, the odds of belonging to the A-level subset were about five times higher in the experimental group (odds ratio = 5.08, Fisher $p < 0.001$).

The choice of writing as the anchor matters for interpretation. In this pilot, writing was based on open-ended performance rather than on a more easily inflated response format, so it served as a comparatively stable reference point for judging whether reading movement was broadly plausible within the local classroom context. The same empirical rule was then applied to both groups without group-specific tuning. This design does not remove all uncertainty, but it makes the filtering logic transparent, reproducible, and less dependent on post hoc threshold selection.

Table 6 compares mean learning gains in the full student sample and in the A-level high-confidence subset. In the full sample, the control group showed a much larger early-stage reading gain ($\Delta R_1 = 19.49$) than the experimental group ($\Delta R_1 = 4.63$, $p < 0.001$), but it then showed a negative later-stage change ($\Delta R_2 = -8.28$), whereas the experimental

Table 4

Conditional reading-score percentiles by mid-term writing-score stratum. n denotes the experimental calibration sample; empty strata are omitted.

Writing-score stratum	n	Q_{15}	Q_{85}
[0, 60)	10	35.98	70.37
[60, 70)	70	36.44	79.42
[70, 80)	64	52.65	86.72
[80, 90)	4	77.18	91.32

Table 5

Confidence-level distribution after applying the data-referenced filtering rule.

Group	A (high)	B (partial)	C (low)	Total
Experimental	107 (72.8%)	33 (22.4%)	7 (4.8%)	147
Control	50 (34.5%)	74 (51.0%)	21 (14.5%)	145

Table 6

Mean learning gains in the full sample and the A-level high-confidence subset.

Group	Subset	ΔR_1	ΔR_2	ΔW
Experimental	Full sample	4.625	7.674	-0.127
Control	Full sample	19.494	-8.282	1.589
Experimental	A-level subset	1.955	10.519	-0.256
Control	A-level subset	4.178	3.508	-1.361

group remained positive ($\Delta R_2 = 7.67$, $p < 0.001$). After the analysis was restricted to A-level cases, the apparent early-stage advantage of the control group disappeared (1.95 vs. 4.18, $p = 0.460$), while the later-stage reading contrast remained and favored the experimental group (10.52 vs. 3.51, $p = 0.0029$, $g = 0.49$). Panel B of Figure 4 shows the same shift visually: once potentially inconsistent rows are removed, the groups begin from more comparable mid-term levels, but the experimental group retains the stronger later-stage rise. Writing gains were unstable across filtering choices: the full-sample contrast was weak ($p = 0.0467$), the stricter valid-confidence comparison became marginal ($p = 0.0518$), and the A-level contrast was no longer significant ($p = 0.346$). The writing results should therefore be treated as exploratory rather than as a robust group difference.

Read together, the confidence distribution and the gain comparison suggest that the issue is not simply whether one group scored higher at one stage. The more important difference is that the experimental group’s trajectory remains comparatively coherent after a stricter filter is imposed, whereas part of the control group’s apparent early advantage is carried by rows that do not survive the same confidence condition. This does not convert the pilot into a causal test, but it does support the narrower claim that governed records and confidence-aware interpretation yield a more stable description of classroom development than raw stage means alone.

The quantitative patterns become more interpretable when read together with classroom use. Platform access records and classroom observations suggested that many reports were viewed soon after publication and that some students revisited them more than once, paying attention not only to the overall score but also to profile views, trend displays, and structured feedback. This suggests that the platform functioned as more than a score-release interface; published outputs also served as reference points for later review and reflection.

From the instructional side, the structured reports appeared to support quicker inspection of borderline or ambiguous cases. Evidence excerpts, dimension-level summaries, and organized result presentation helped teachers conduct sampling-based quality control without reassessing every output from scratch. When needed, regeneration or re-evaluation could be triggered before publication. These observations remain informal rather than instrument-based evidence, but they are consistent with the claim that the governed workflow supported teacher review rather than bypassing it.

Overall, the pilot supports a restrained interpretation. The method can produce reviewable structured outputs, preserve ensemble metadata, and support a classroom workflow in which teachers retain control over publication. It does not establish causal instructional superiority, but it does show that a confidence-aware reading of governed assessment data yields a more stable and more interpretable picture than raw score change alone.

5. Discussion

The main contribution of this study is not autonomous scoring, but a governed workflow for making variable model generations usable in classroom assessment. Much of the existing literature asks whether LLM outputs approximate human scoring under controlled conditions. The present results suggest that classroom deployment raises a different requirement: outputs must become publication-ready records that can be validated, normalized, stabilized, reviewed, and later reused. From that standpoint, the JSON-only contract, deterministic recomputation, traceability metadata, and teacher checkpoint are not auxiliary safeguards around the method; they are the method’s core mechanism for converting generation into accountable assessment artifacts.

That point also clarifies why structured output alone is not enough. A model may produce a seemingly well-formed assessment object while still varying in field completeness, score consistency, or evidential emphasis across repeated calls. In a classroom platform, those objects are not ephemeral generations; they feed report pages, trend views, and later teacher follow-up. Deterministic post-processing and metadata retention therefore serve a practical educational purpose: they transform a variable completion into a record that can be inspected, compared, regenerated, and, if necessary, withheld from publication until a teacher is satisfied with the result.

A further implication is that classroom assessment should not be modeled as a single inference event. In ordinary teaching, evaluation involves release timing, selective review, borderline-case inspection, and follow-up explanation after publication. If generation, explanation, and release are collapsed into one model call, the system becomes hard to audit and even harder to use responsibly when outputs are inconsistent. By contrast, the governed workflow makes these stages explicit. It separates upstream model variability from downstream pedagogical release, which is precisely what allows teacher judgement to remain visible rather than being hidden behind a final number.

The pilot results also illustrate why confidence-aware interpretation is useful in real instructional environments. Classroom assessments do not occur under perfectly uniform conditions, so raw score movement may combine learning change with administration noise, irregular participation, or transient performance variation. The filtering rule introduced here was designed to reduce that interpretive risk at the group level rather than to label individual students. The contrast between the full sample and the A-level subset suggests that the experimental group’s later-stage reading pattern is more robust to the removal of potentially unstable observations, whereas part of the control group’s apparent early advantage is more sensitive to filtering. In that sense, confidence-aware analysis plays the same role in the evidence layer that validation and stabilization play in the runtime layer: it reduces over-interpretation of variable outputs. At the same time, the pilot does not directly validate the four competence-dimension scores as standalone psychometric scales; it tests whether assignment evidence can be converted into governed, reviewable records that remain interpretable enough for teacher-mediated classroom use.

This point matters not only for one pilot comparison, but also for longitudinal use. Classroom reports are often revisited across tasks, displayed in profile views, or used to support later teacher-student conversations. Once results enter that longer cycle, instability at one stage can distort later interpretation even if the original score looked plausible in isolation. The combination of stabilized outputs and confidence-aware analysis therefore supports a narrower but educationally meaningful aim: preserving enough comparability that later discussion can focus on instructional response rather than on whether the underlying record is itself too unstable to trust.

Several boundaries remain important. The pilot was conducted in a single-course setting, relied mainly on open-ended written tasks, and used informal classroom observations rather than standardized feedback instruments. The workflow is therefore better understood as a practical, exploratory classroom method than as a completed causal evaluation. The confidence filter also depends on an experimental-group reference distribution anchored by mid-term writing, and the present paper does not report sensitivity checks with alternative reference groups, anchors, or threshold settings. More broadly, the current design is best suited to tasks whose evidence can be meaningfully

expressed in text, even though the surrounding platform also includes behavior insight, speaking practice, and goal tracking. Those surrounding modules provide system context rather than the primary empirical validation target of this paper. Teacher review likewise remains indispensable, which means the system should be understood as a human-supervised support layer rather than as a substitute for teacher judgement [14].

This emphasis on teacher mediation is consistent with formative-assessment theory. In classroom practice, feedback is most useful when it helps teachers and students organize subsequent action rather than merely attach a score to completed work. The platform evidence reported here suggests that structured reports, trend displays, and dimension-level feedback can support that process by giving teachers a faster way to inspect ambiguous cases and by giving students a more interpretable basis for later reflection. The contribution is therefore not only technical robustness, but also a workflow shape that fits more naturally with co-regulated classroom use than with unattended automation.

These limits also suggest concrete next steps. Future validation should compare single-pass outputs with stabilized outputs on the same assignment sets, study when teachers trigger regeneration or override publication, and test whether the same governance logic transfers to spoken interaction and other richer evidence forms. Such extensions would make it possible to estimate more directly how much of the practical value comes from the governance layer itself and how far the current workflow generalizes beyond one course context.

6. Conclusion

This study presents a governed workflow for LLM-based formative assessment that converts variable model generations into stable classroom records through competence-oriented task decomposition, a JSON-only output contract, normalization, score stabilization, deterministic post-processing, and teacher review. Technical verification showed that the pipeline can run end to end in the verified deployment environment. The classroom pilot further showed that confidence distributions differed markedly between groups and that, within the A-level high-confidence subset, the experimental group displayed a stronger later-stage reading gain than the control group.

These findings are bounded by the single-course setting and the exploratory design, and they do not by themselves validate the four competence-dimension scores as standalone psychometric measures or justify broad causal claims about learning improvement. Even so, they indicate that governed processing can make LLM-supported assessment more reviewable and more usable in routine teaching. Future work should test the contribution of the governance layer more directly, examine teacher-review behavior in greater detail, and extend the same logic to spoken and other richer forms of classroom evidence.

Acknowledgment

The authors declare that there is no external funding and no conflict of interest.

References

- [1] Core Competency Research Group, Core Competencies for Chinese Students' Development, Journal of The Chinese Society of Education, 2016, vol. 0, no. 10, pp. 1–3.
- [2] OECD, OECD Learning Compass 2030: A Series of Concept Notes, 2019. URL: https://www.oecd.org/content/dam/oecd/en/about/projects/edu/education-2040/1-1-learning-compass/OECD_Learning_Compass_2030_Concept_Note_Series.pdf.
- [3] X. Wang, Personalized forgetting-aware multi-view graph knowledge tracing for vocational bachelor's practical training platform evaluating, IFS/ACM Transactions on Machine Learning, 2026, vol. 3, no. 1, pp. 20–29.
- [4] X. Tang, H. Chen, D. Lin, K. Li, Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments, Heliyon, 2024, vol. 10, no. 14, pp. e34262:1–e34262:18.
- [5] D. W. Zhang, M. Boey, Y. Y. Tan, A. H. S. Jia, Evaluating large language models for criterion-based grading from agreement to consistency, npj Science of Learning, 2024, vol. 9, pp. 79:1–79:4.
- [6] A. Pack, A. Barrett, J. Escalante, Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability, Computers and Education: Artificial Intelligence, 2024, vol. 6, pp. 100234:1–100234:9.
- [7] W. Li, H. Liu, Applying large language models for automated essay scoring for non-native Japanese, Humanities and Social Sciences Communications, 2024, vol. 11, pp. 723:1–723:15.
- [8] T. Wang, S. Yin, Enhancing sustainability of online education via a validated cloud management framework, IFS/ACM Transactions on Machine Learning, 2026, vol. 3, no. 1, pp. 9–19.
- [9] C. Grévisse, LLM-based automatic short answer grading in undergraduate medical education, BMC Medical Education, 2024, vol. 24, no. 1, pp. 1060:1–1060:16.
- [10] Z. Chen, T. Wan, Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy, Physical Review Physics Education Research, 2025, vol. 21, no. 1, pp. 010126:1–010126:25.
- [11] L. Rodrigues, C. Xavier, N. Costa, D. Gasevic, R. F. Mello, Is GPT-4 fair? An empirical analysis in automatic short answer grading, Computers and Education: Artificial Intelligence, 2025, vol. 8, pp. 100428:1–100428:12.
- [12] Y. Wang, J. Huang, L. Du, Y. Guo, Y. Liu, R. Wang, Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity, Computers and Education: Artificial Intelligence, 2025, vol. 9, pp. 100481:1–100481:15.
- [13] P. Baldwin, An audit-style framework for evaluating bias in large language models, Frontiers in Education, 2025, vol. 10, pp. 1592037:1–1592037:9.
- [14] M. J. Veugen, J. T. M. Gulikers, P. den Brok, Secondary school teachers' use of formative assessment practice to create co-regulated learning, Journal of Formative Design in Learning, 2024, vol. 8, pp. 15–32.
- [15] E. Panadero, A review of self-regulated learning: Six models and four directions for research, Frontiers in Psychology, 2017, vol. 8, pp. 422:1–422:28.
- [16] P. Prasad, A. Sane, A self-regulated learning framework using generative AI and its application in CS educational intervention design, Proceedings of the 55th ACM Technical Symposium on Computer Science Education, ACM, Portland, OR, USA, 2024, pp. 1070–1076.
- [17] Z. Chu, S. Wang, J. Xie, T. Zhu, Y. Yan, J. Ye, A. Zhong, X. Hu, J. Liang, P. S. Yu, Q. Wen, LLM agents for education: Advances and applications, Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 13782–13810.
- [18] A. T. Neumann, Y. Yin, S. Sowe, S. Decker, M. Jarke, An LLM-driven chatbot in higher education for databases and information systems, IEEE Transactions on Education, 2025, vol. 68, no. 1, pp. 103–116.
- [19] T. Bowles, J. Scull, S. R. Jimerson, School connection through engagement associated with grade scores and emotions of adolescents: Four factors to build engagement in schools, Social Psychology of Education, 2022, vol. 25, no. 2-3, pp. 675–696.
- [20] A. Arvatz, R. Peretz, Y. J. Dori, Self-regulated learning and reflection: A tool for teachers and students, Metacognition and Learning, 2025, vol. 20, pp. 15:1–15:24.
- [21] F. Stebner, C. Schuster, X. L. Weber, S. Greiff, D. Leutner, J. Wirth, Transfer of metacognitive skills in self-regulated learning: Effects on strategy application and content knowledge acquisition, Metacognition and Learning, 2022, vol. 17, no. 3, pp. 715–744.
- [22] A. Landowska, K. Budzyska, H. Zhang, Quantitative and qualitative analysis of moral foundations in argumentation, Argumentation, 2024, vol. 38, no. 3, pp. 405–434.