Evaluation of Student Assignments Using Local Large Language Models and Graph-Based Evidence

Neven Pintarić
University of Zadar
Department of Information Sciences and Technologies
Ul. dr. F. Tuđmana 24i, Zadar, Croatia
neven@unizd.hr

Ivan Peraić
University of Zadar
Department of Applied Communication Studies
Ul. dr. F. Tuđmana 24i, Zadar, Croatia
iperaic@unizd.hr

Abstract. *The paper presents an experimental system, Graph2LLM, for evidence-based graph evaluation of student term papers using a locally implemented LLM. The system first deterministically maps paper content into a graph with linked evidence and then uses the local LLM to validate the assignment requirement nodes. Evaluation of 25 student papers across two repeated runs shows that the deterministic mapper identifies 132 out of 150 AR4–AR9 instances. Through the local LLM, 129 instances were validated per run. The experimental results demonstrate 94.57% agreement of LLM decisions, 98.45% agreement on evidence sufficiency, and an average confidence difference of 0.0229. A human evaluation of all 150 AR instances was carried out by two evaluators. Differences were resolved by consensus. The comparison shows that the system is precise where it affirms full coverage of requirement, but stricter than human evaluators elsewhere.*

Keywords. automated essay scoring, evidence-based evaluation, graph-based evaluation, DigComp, local LLM

1 Introduction

Large language models (LLMs) have become relevant tools in the field of education, especially in the analysis and generation of textual content and in the automated evaluation of student papers. Research (Song et al., 2024) points to the potential of LLMs in the evaluation of student papers, while research (Narreddy, Joordens, & Prompiengchai, 2025) shows the possibility of integrating them into different educational contexts, including automatic assessments conducted by teachers.

The classic approach in which the student paper is sent directly to the language model reveals several methodological limitations documented in recent research: inconsistencies and variability of outputs (Pack, Barrett, & Escalante, 2024), reduced transparency and inability to revise evaluation decisions (Kumar & Boulanger, 2021), and inconsistent application of criteria due to sensitivity to query formulation (prompt sensitivity problem).

Taken together, these limitations suggest that direct LLM evaluation can reduce the repeatability, verifiability and consistency, which is problematic in educational contexts where evaluation must be methodologically grounded and equally applicable to all students.

A further limitation relates to transparency. In direct evaluation using an LLM, it is often unclear which parts of the student paper support a model decision, and it is not easy to verify the textual evidence on which assessment is based (Xu et al., 2026). This reduces the auditability and reliability of the system, which is one of the key issues in educational contexts.

Existing systems are mainly focused on generating a final assessment or certain aspects of feedback in isolation, while structured and integrated feedback that includes the coverage of task requirements and guidelines for improvement is often absent (Narreddy, Joordens, & Prompiengchai, 2025). In practice, students receive little information about which parts of their work are missing, which task requirements are met, and how the quality of their work can be improved (Li et al., 2026). The absence of these elements reduces the potential of LLMs as learning support systems.

These problems are further reinforced when evaluating a larger set of seminar papers in the same field. In such cases, it is expected that the system produces stable and comparable results for input that is identical or similar in content, without significant differences between individual evaluations, while maintaining a uniform criterion for comparing student papers.

Student work may contain personal data: information that can identify an individual, directly or indirectly. Sending such content to external LLM services therefore raises questions not only of compliance with data protection, but also of intellectual property. A locally implemented LLM addresses both of these challenges by ensuring that all data remains within the institutional infrastructure, without being transferred to third-party services (Zesch et al., 2024).

To address these limitations, this paper introduces Graph2LLM, an experimental system for evidence-based graph evaluation of student papers using a locally implemented LLM. The local LLM does not serve as a primary evaluator, but as a part of a structured process that includes the construction of a reference graph, deterministic content mapping, and evidence-based node validation. In its experimental phase, Graph2LLM is a teacher support and quality control instrument for verifying the coverage of assignment requirements (AR); it does not produce grades or scores for student evaluation.

The paper makes five contributions: a.) Graph2LLM architecture that separates deterministic mapping from LLM validation, b.) an evidence-based evaluation approach built on a graph representation, c.) empirical verification of reproducibility, d.) evaluation on 25 student term papers (STP) through two repeated runs with a locally implemented LLM and e.) a human evaluation of all 150 AR instances.

Building on these contributions, this work addresses the following research questions:

RQ1: Does the deterministic graph-based mapper produce consistent and rule-governed AR node identification across all student seminar papers when applied repeatedly under identical conditions, regardless of input document quality?

RQ2: Does a locally implemented large language model produce consistent validation decisions for assignment requirement nodes when the same evidence-based payload is sent repeatedly under identical conditions, and does the AR5 decision remain equally stable when it is additionally validated against the DigComp 3.0 reference context?

RQ3: Can an evidence-based approach built on a graph representation distinguish between fully covered, partially covered, missing, uncertain, and potentially wrongly mapped assignment requirement nodes?

RQ4: On which AR nodes, and for what reasons, do the decisions of the proposed architecture differ from the consensus decisions of human evaluators?

This study is presented as follows. Section 2 reviews related work on LLM-based paper evaluation, automated essay scoring, rubric-based approaches, and graph-based assessment. Section 3 describes the architecture and methodology. The experimental setup and results are presented in Section 4. The findings and limitations are discussed in Section 5, and the conclusions are given in Section 6.

2 Literature Review

2.1 LLM in Evaluation of Student Papers

The development of LLMs has opened new possibilities for applying artificial intelligence in educational assessment, in evaluating student essays and other written work. Unlike traditional systems that rely on predefined linguistic features, modern LLMs enable analysis of the full text and consider the semantic, argumentative and structural characteristics of written work (Pack, Barrett, & Escalante, 2024).

Existing research shows that LLMs can achieve a substantial level of agreement with the assessments of human evaluators. For example, Ishida et al. (2024) found that using clearly defined rubrics and pairwise comparison of papers results in high correlation between model and teacher assessments. Similarly, Liu, Ye, & Yan (2026) show that some modern models can achieve a high level of consistency and alignment with expert raters when evaluating student essays.

However, research also points to methodological limitations. Among the most frequently noted problems are variability of results across repeated runs, sensitivity to prompt formulation, and differences in the interpretation of evaluation criteria (Huang et al., 2025; Pack, Barrett, & Escalante, 2024). In addition, the evaluation decisions generated by LLMs are often characterized by limited transparency, since it is not always possible to determine which parts of the student paper form the basis for a particular assessment (Kostic et al., 2024). Chiang et al., (2024) also warn that models do not always consistently follow the given evaluation criteria and can be susceptible to manipulation through prompt design.

These findings indicate that, despite the substantial potential of LLMs in educational assessment, questions of reliability, transparency and verifiability of evaluation decisions remain open. For this reason, recent research focuses on structured and evidence-based approaches that link evaluation decisions to explicit textual evidence and enable their subsequent verification and revision.

2.2 Automated Essay Scoring

Automated Essay Scoring (AES) refers to the use of computational methods for the automatic evaluation of written work. Early AES systems were based on statistical models and surface linguistic features such as text length, word frequency and grammatical correctness, with systems such as PEG and e-rater laying the foundations for automated assessment in education (Elmassry et al., 2025).

Advances in machine learning and deep learning enabled a shift from manually defined features to models that autonomously learn text representations. The use of neural networks, transformers and pretrained models such as BERT improved the systems' ability to recognize semantic and discourse patterns and increased agreement with human evaluators (Elmassry et al., 2025; Mughal et al., 2026). At the same time, AES systems have shown substantial potential to reduce grading time and increase the consistency of assessment, in settings with a large number of student papers (García-Varela et al., 2025). Despite this progress, AES systems still face limitations. Models based on surface linguistic features

often fail to adequately assess the quality of argumentation, textual coherence, creativity and the relevance of content to the assigned task (Atkinson & Palma, 2025). In addition, numerous studies warn of problems with generalization across domains, sensitivity to scoring criteria, and challenges related to fairness and explainability of models (Elmassry et al., 2025; Mughal et al., 2026; García-Varela et al., 2025).

Although modern models achieve better results compared to human raters, questions of transparency and verifiability of evaluation decisions remain open research challenges.

The problem of explicitly linking the assigned score to concrete textual evidence that supports the evaluation decision is pronounced. For this reason, recent research turns toward approaches that enable transparent tracking of the evaluation process and the subsequent verification and revision of automatically generated assessments (Xu et al., 2026).

2.3 Rubric-Based Evaluation

For the structured evaluation of student work, rubrics have been most often used as tools lately. The rubrics are based on predefined criteria through which different aspects of the quality of work, such as content, organization, and language quality, are evaluated separately (Yavuz, Çelik & Yavaş, 2025).

The evaluation process with this method is more transparent and allows students to gain insight into the reasoning behind the grade.

Their further application can be seen in order to reduce the subjectivity of evaluation and increase consistency among evaluators. Furthermore, research shows that rubrics not only serve as an assessment tool, but also as a mechanism for providing structured feedback, thus expanding their contribution to the development of writing skills as well as self-regulated learning (Alghizzi and Alshahrani, 2024).

This approach is also applied in modern automated evaluation systems, which integrate rubrics into the evaluation process to improve the reliability and explainability of the results.

Research (Yavuz et al., 2025) shows that large language models can achieve a high level of agreement with human evaluators when evaluating work according to predefined analytical criteria. Similarly, Vangelova and Gancheva (2026) propose a framework that combines large language models, Retrieval-Augmented Generation (RAG), and analytic rubrics so that evaluation decisions are based on clearly defined criteria and contextual evidence.

While such systems show positive aspects, rubric-based systems often face challenges related to the interpretation of criteria and the justification of the grades awarded. Modern models achieve high accuracy but often do not provide a sufficient explanation of how individual criteria influenced the final estimate (Kumar and Boulanger, 2021).

2.4 Graph-Based Approaches and Evidence-Based Assessment

Graphs provide a framework for modelling relationships between entities, and knowledge graphs have been applied across various domains (Ji et al., 2022). In education, graph-based approaches model learning outcomes and educational structures, giving an organised representation of concepts and their interconnections that supports content recommendation, learning-outcome prediction, and educational evaluation (Qu et al., 2024). Entities are represented as nodes and their relationships as edges, which makes the underlying data interpretable (Li et al., 2024).

Progress in LLM has motivated research into combining them with graph structures, so as to join the semantic understanding of text with an explicit representation of structural relationships in the data. Such a combination links the implicit knowledge held in model parameters with explicitly represented facts, which improves interpretability and reasoning as well as the reliability of generated responses (Li et al., 2024; Pan et al., 2023). LLMs interpret textual content effectively, whereas graph structures expose relationships that often remain implicit in the text. GraphRAG illustrates this: it uses a graph-based knowledge representation to organise and summarise information from large text corpora and to analyse related concepts and entities (Edge et al., 2024).

In educational measurement, evidence-based assessment has become important. Evidence-Centered Design (ECD) is a prominent example, grounding assessment decisions in an explicit chain of reasoning that links student responses to identified evidence and to the constructs being assessed (Arieli-Attali et al., 2019). This contributes to the validity, transparency, and interpretability of the assessment process. ECD has recently been applied to LLM-based conversational assessment, where it structures prompts so that the model elicits and records evidence of student knowledge (Zapata-Rivera et al., 2024).

These principles also extend to graph-based approaches, where graphs do more than represent knowledge: they also guide reasoning and support factchecking and validation of LLM output (Ma et al., 2025). Rubric-based assessment, explainable artificial intelligence, and graph-based knowledge representation have all advanced, yet integrating AR, student content, and explicit textual evidence within a single evaluation framework remains an active research problem.

3 Methodology

3.1 System Overview

The proposed methodology evaluates STP through a graph-based, evidence-based validation process. The

system is implemented as a two-node distributed architecture. The orchestration layer runs on a computer in a combination of Windows, Python, Django with a PostgreSQL database. A local LLM validation layer was established on a MacBook Pro M4 computer with 48 GB RAM through LM Studio (0.4.12) with the locally implemented model *qwen3.5-35b-a3b-uncensored-hauhaucs-aggressive* (GGUF, Q5_K_M). This variant was chosen deliberately. An uncensored LLM is one whose safety training has been removed, and this removal is known to weaken the general capability of the model and to make its output less predictable (Qi et al., 2024; Young, 2026). It is therefore the hardest case for a system that must produce repeatable decisions. The design intention is that stability should be achieved through the architecture described in the following sections, and not through the LLM model itself. Communication between nodes and layers takes place over a local network through an OpenAI-compatible REST API endpoint.

Rather than directly evaluating the full text of a student term paper (STP) using a local LLM, Graph2LLM applies a structured four-layer processing pipeline, illustrated in Fig. 1.

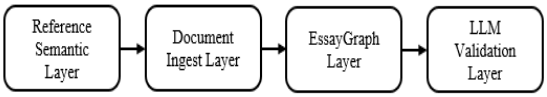


Figure 1. Graph2LLM processing pipeline

The reference semantic layer, comprising the DigComp Graph and the Assignment Graph, is constructed in advance and serves as the foundation for all subsequent processing. The system then converts each STP into a structural graph, deterministically maps it against predefined AR, and only then validates individual requirement nodes using the local LLM on the basis of evidence segments linked to each node. Each layer is described in detail in the following subsections.

3.2 Formal Graph Layers

Three formal graph layers were defined: DigComp Graph, Assignment Graph and Essay Graph.

Serving as reference model of digital competences, the DigComp Graph is modelled on the basis of the publicly available JSON-LD dataset for DigComp 3.0 framework (Directorate-General for Employment, Social Affairs and Inclusion, 2025). It contains competence areas, competences and learning outcomes with associated properties such as outcome types and proficiency levels. In this layer, the node is the primary structural unit of the graph, carrying learning outcomes as properties that serve as the evidential basis for validation. Learning outcomes represent a measurable and operationalizable foundation for automated assessment (Divjak, Svetec, & Horvat, 2024). It is used exclusively where the task requirement directly asks the student to identify and describe a DigComp area, competences and learning outcomes.

Capturing the pedagogical model of the STP; the Assignment Graph defines the expected requirements through ten requirement nodes (AR1–AR10). These requirements include formal document elements, title page (AR1), table of contents (AR2) and references (AR10), and substantive elements, including an introduction (AR3), a definition of digital competence (AR4), a selected DigComp area (AR5), a self-assessment (AR6), relevance to early childhood education practice (AR7), a personal development plan (AR8) and a conclusion (AR9).

The Essay Graph is created from the provided STP. Segment nodes preserve the order of appearance in the document, ensuring both structural integrity and traceability to the original text. Any subsequent validation decision must be linked to concrete evidence from the STP.

3.3 Document Ingest Layer

The evaluation process begins with uploading an STP document that is stored as a document object with a basic metadata. The ingest stage extracts document pages (*DocumentPage*), text segments (*DocumentSegment*), and evidence records (*Evidence*). This phase is deterministic and does not involve a local LLM. After the ingest phase, an initial structural *EssayGraph* is built in which each relevant document segment becomes an *EssayNode*, while the relationships within the document are represented as *EssayEdge* objects.

3.4 Deterministic Assignment Mapping

This step constructs the semantic assignment layer through a deterministic mapper that implements an architectural principle, considering prior work showing that evidence components must be explicitly identified and structured before LLM scoring (Wang et al., 2026). Unlike LLM-based extraction approaches, mapping in this work is implemented deterministically through configurable *AssignmentRequirementProfile* records. Each profile describes a single task requirement and contains parameters such as requirement code, name, requirement type, expected order, title aliases, negative patterns, expected evidence types, evidence policy, visible support segment limits, and acceptance thresholds.

The system groups document segments into sectional units and compares each sectional group with the active requirement profiles. Structural segmentation in unstructured documents according to section headings with variable formats represents an NLP problem (Tannous, Ramadan, Rajab, 2023).

The result of the mapping is calculated by combining multiple signals and heuristic value: section title similarity (0.34), structural order (0.20), section

number (0.14), evidence type compatibility (0.10), preface bonus (0.08), content keyword match (0.14), negative patterns, and TOC penalty. These signals are aggregated into a weighted mapping score with fixed weights calibrated on baseline documents. Each candidate mapping is then classified as accepted, rejected, or flagged for review. Only accepted candidates become semantic AR nodes in *EssayGraph*.

As an input control before local LLM validation, a baseline graph diagnostic is performed, which checks for the presence of expected AR nodes, missing nodes and suspicious mappings.

Deterministic mapping is reproducible under the same input conditions: the same document, the same version of the ingest procedure, the same configuration of the *AssignmentRequirementProfile* records, and the same version of the code produce the same result. Consistency in this context does not necessarily imply correctness, but rather the repeatability and auditability of mapping decisions. This is particularly important in the educational context, where transparency, verifiability and ex-post auditability of automated evaluation decisions are considered methodological requirements (Kim et al., 2025).

3.5 Assignment Requirement

Each AR node belongs to one of two methodological categories. Structural requirements describe formal document properties and include AR1, AR2 and AR10. Content requirements are mapped to actual textual sections of the paper and include AR3, AR4, AR5, AR6, AR7, AR8 and AR9. Although AR3 belongs to the content requirements, it is excluded from the initial local LLM validation set because introductions are generally of a general nature and do not contain a specific DigComp validation structure. Accordingly, local LLM validation in this paper is focused on AR4–AR9.

3.6 Evidence-Based Local LLM Validation

LLM-based validation has known limitations (Zheng et al., 2023). This can be mitigated by clearly defined multi-dimensional rubrics, in which the model does not produce a holistic assessment but instead responds to pre-defined evaluation questions (Hashemi et al., 2024).

After deterministic mapping, the local LLM validation layer validates identified individual AR nodes with explicitly linked evidence segments. The model does not receive the entire STP, but rather a structured payload for a single AR node containing document metadata, graph identifier, assignment node code and name, requirement description, section headings, deterministic mapper confidence, and evidence texts linked to that AR node. In this way, validation is limited to the verification of a clearly defined requirement and the associated evidence.

This logic is operationalised in the local LLM validator prompt. The prompt defines the model as a JSON-only validation API and explicitly requires the response to be a single valid RFC8259 JSON object with no additional text, comments, markdown or visible reasoning. Furthermore, the prompt instructs the model to use only *the submitted evidence_texts* and the reference context provided, without using outside knowledge, fabricating facts, or assuming that the missing content exists elsewhere in the document.

Every response from the local LLM must return a structured JSON response validated against a pre-defined schema. It includes the assignment node code, decision, evidence sufficiency, detected requirement, confidence, problems, reasoning summary and an object for DigComp reference validation. Permitted decisions are *fully_covered*, *partially_covered*, *missing*, *wrong_mapping* and *uncertain*, interpreted under a minimum evidence-sufficiency threshold: a node with insufficient semantic content is *marked* missing with *evidence_sufficient=false* rather than *partially_covered*.

For content requirements AR6, AR7 and AR9, a stricter interpretation rule is applied: if the evidence is too short, generic, poorly reasoned or does not contain requirement-specific semantic content, the decision must be *missing* rather than *partially_covered*.

A further part of the prompt distinguishes between assignment coverage validation and DigComp reference alignment validation. If *digcomp_reference_context* is not provided, the model must not perform exact alignment with DigComp areas, competences or learning outcomes, but validates only the coverage of the task requirement. If *digcomp_reference_context* is provided, the model may use it exclusively as the supplied reference context and must not fabricate DigComp codes, areas, competences, learning outcomes or proficiency levels.

A *wrong_mapping* decision marks cases where deterministic mapper accepted candidate mapping but the local LLM judges, from the linked evidence, that it does not correspond to the assigned AR node, separating the formal presence of section from the semantic correctness of its evidence.

3.7 Assignment Coverage

Assignment coverage validation is applied to all semantic nodes included in the initial validation set, i.e. AR4–AR9. The coverage criteria differ by requirement: a defined digital competence (AR4); an explicit self-assessment with meaningful reasoning about the proficiency level and a relevant example (AR6); a concrete pedagogical application connecting the selected area or competence to the work of the early childhood educator (AR7); a development plan with a goal, a current or desired proficiency level, and a means of achievement, with partial coverage permitted only when at least two mandatory plan elements are present (AR8); and a conclusion synthesising the topic

of the paper, the selected area or competence, and the main findings (AR9).

Node AR5 is a special case, as it directly links the AR to the DigComp framework. Therefore, AR5 is validated both against the Assignment Graph and against the DigComp Graph through DigComp reference alignment validation. This allows the model to verify whether the student's description is semantically aligned with the detected DigComp area, competences and learning outcome dimensions. In the current phase, AR5 is the only node with dual validation. The results are not interpreted as a complete DigComp validation of the entire student paper.

3.8 Architecture Evaluation Metrics

The evaluation is based on node-level graph and validation metrics aligned with the research questions RQ1 to RQ3. For the deterministic mapper, we measure repeated-run agreement, AR coverage rate, missing node count and wrong mapping count. For local LLM validation, we measure decision distribution, evidence sufficiency rate, confidence score distribution and decision stability across repeated runs. Node AR5 is reported separately because it is the only node with additional DigComp Graph alignment, while AR4, AR6, AR7, AR8 and AR9 are validated only against the Assignment Graph. For each run, the system stores the AR node, evidence count, decision, confidence, prompt version, model metadata, run identifier and DigComp context usage. This makes it possible to verify the factual basis for each validation decision at a later stage.

3.9 Human Evaluation Protocol

Once the architecture had shown stable and repeatable decisions across the two runs, the next step was to test whether those decisions were also correct. Human evaluation was carried out in order to establish a reference against which the correctness of the system's decisions, and not only their reproducibility, could be assessed. Two evaluators independently coded all 150 AR instances (25 STPs × AR4–AR9). The evaluators are the authors of this research. To limit the resulting bias, the following procedure was applied: coding was blind, so that neither evaluator saw the system output or the decisions of the other, and the papers were presented in a randomised but identical order. The evaluators followed a written set of coding rules that mirrored those applied by the system, namely the minimum-sufficiency threshold, the stricter interpretation of AR6, AR7 and AR9, and the two-of-three rule for AR8. Decisions were based exclusively on the text of the paper, mirroring the constraint imposed on the local LLM by the prompt. Instances on which the two evaluators initially disagreed were resolved by consensus, in which the relevant text segment was re-read, the written rules re-examined, and the final decision reached through reasoned argument. A measurement instrument with two parts was prepared, one for each layer of the architecture. Part A records whether content addressing the requirement is present at all, irrespective of quality, and serves as the reference for the deterministic mapper. Part B assigns one of the five decision categories used by the system and serves as the reference for local LLM validation. Inter-rater agreement was computed on the independent codings; the disagreements were then resolved by consensus between the two evaluators; the reported kappa values are computed on the independent codings and are therefore not inflated by the consensus step. For the deterministic mapper, Part A provides precision, recall and F1 against the consensus presence decisions. For the local LLM validation layer, agreement with the consensus decisions in Part B is reported as a confusion matrix of system against human decisions, percentage agreement, Cohen's κ, and, for the *fully_covered* category, precision, recall and F1.

3.10 Ethics

This experimental research was conducted in accordance with the policy on the use of artificial intelligence at University of Zadar (University of Zadar, 2026), which recommends locally implemented solutions for sensitive data and prohibits automated decision making without human oversight. Approval was obtained from the Ethics Committee of the Department of Information Sciences and Technologies for the implementation of the research (decision of 10 July 2026, ref. 602-03/26-04, no. 2198-1-79-33-26-01). All processing was carried out locally on institutional infrastructure, and only aggregated results are reported, so that no individual STP can be identified. The experiment was run after the winter semester and all examination has been completed, and it had no influence on scoring or the final grades of the students enrolled in the course. Students were informed that their term papers could be used for research purposes.

4 Results

4.1 Experimental Set-up

Experimental evaluation is carried out on a set of 40 STPs from the winter course "Child and Computer", taught in the winter semester of the 2025/2026 academic year in the undergraduate university study programme in Early and Preschool Education at University of Zadar. The course includes the preparation of a seminar term paper on digital competencies according to the DigComp framework. We have three subsets of STP papers. For the purpose of development, 5 papers were used. System calibration and baseline testing were performed on 10 papers, and 25 papers were used for final testing.

Each STP goes through two independent validation runs (*run_01* and *run_02*) under identical conditions: the same LLM, temperature (0.1), prompt version (*assignment_node_validation_v1*), context window 16,384 tokens and *Assignment-RequirementProfile* configuration.

4.2 Baseline Testing and Calibration

Baseline testing was conducted on 10 STPs using the locked *AssignmentRequirementProfile* configurations for AR4–AR9. The deterministic mapper identified 49 of 60 expected AR nodes (81.67% coverage). Repeated runs produced identical mapper output, confirming deterministic reproducibility.

Local LLM validation revealed that table content in AR8 (personal development plan) was not consistently extracted as separate evidence segments. In these cases, the local LLM validated only the chapter heading rather than the table content. This resulted in systematic misclassification.

To address this, the ingest layer was extended to extract table rows as individual evidence segments. Prompt calibration was subsequently performed on a purpose-built control document containing formally correct but semantically thin chapters alongside a properly structured AR8 table. This control document was designed to verify that the local LLM distinguishes formal chapter presence from semantically sufficient coverage. After calibration, the local LLM correctly classified generic single-sentence evidence as *missing* and complete development plan tables as *fully_covered*.

All mapping weights, acceptance thresholds, negative patterns, *AssignmentRequirementProfile* configurations, prompt settings and LLM inference parameters were derived exclusively from the development and calibration subsets and were locked before the final test set was processed. The 25 final-test STPs were not used in any tuning step, which limits the risk of overfitting.

4.3 Testing Results for Run 1 and Run 2

Results for Run 1 and Run 2 are presented in Table 1. The deterministic mapper processed 150 expected AR4–AR9 instances across 25 documents, identifying 132 instances (88.00%) with identical output in both runs. Of the 21 *Not_validated* instances, 18 correspond to AR nodes not detected by the deterministic mapper, while 3 reflect local LLM processing failures in partial-status validation runs. The 7 recorded decision changes produced a net distributional shift of four instances out of *fully_covered* (38 to 34), appearing as three additional *partially_covered* and one additional *missing*, all between adjacent categories. No transitions between non-adjacent categories were observed.

Table 1. Final testing results Run 1 and Run 2

Metric	Run 1	Run 2
Expected AR4–AR9 instances	150	150
Mapper: found	132	132
Mapper: missing	18	18
Local LLM validated instances	129	129
Decision: fully_covered	38	34
Decision: partially_covered	64	67
Decision: missing	27	28
Not_validated	21	21
Evidence sufficient: True	96	96
Evidence sufficient: False	33	33

Cross-run stability metrics are shown in Table 2. Decision agreement on 129 comparable pairs reached 94.57%, with evidence sufficiency agreement of 98.45%. Seven decision changes and two evidence sufficiency changes were recorded across both runs. The average absolute confidence difference was 0.0229, indicating low numerical variability of the local LLM.

Table 2. Cross-run stability metrics (Run 1 vs Run 2)

Stability Metric	Value
Mapper agreement	150/150 (100%)
Decision agreement (comparable pairs)	122/129 (94.57%)
Evidence sufficiency agreement	127/129 (98.45%)
Decision changes	7
Evidence sufficiency changes	2
Average confidence difference	0.0229

The results per AR node are shown in Tables 3 and 4. Node AR6 and AR9 achieved complete agreement in the decision (100%).

Table 3. AR node decision stability (Run 1 vs Run 2)

AR Node	Comparable Pairs	Decision Agreement
AR4	19/25	18/19 (94.74%)
AR5	22/25	21/22 (95.45%)
AR6	19/25	19/19 (100%)
AR7	24/25	21/24 (87.50%)
AR8	23/25	21/23 (91.30%)
AR9	22/25	22/22 (100%)

The highest variability occurs in AR7 (87.50%) with 3 changes, with problems related to insufficient pedagogical application and incomplete reference context of DigComp. Changes in the sufficiency of evidence occurred only in AR8, in 2 cases, suggesting structural variation in the content of the personal development plan. Node AR5, validated against both

the Assignment Graph and the DigComp 3.0 reference context, showed high stability with a single *fully_covered* to *partially_covered* transition.

All recorded decision changes occurred between adjacent decision categories. This pattern confirms that local LLM variability is concentrated in borderline semantic cases rather than in stable instances of clear coverage or clear absence of evidence.

Table 4. AR node evidence sufficiency and confidence (Run 1 vs Run 2)

AR Node	Evid. Suff. Agreement	Avg. Confidence	Decision Changes
AR4	19/19	0.7737	1
AR5	22/22	0.7695	1
AR6	19/19	0.7563	0
AR7	24/24	0.7458	3
AR8	21/23	0.8443	2
AR9	22/22	0.6773	0

4.4 Human Evaluation Results

This section compares the system output with the human evaluation: the deterministic mapping is examined through Part A of the human coding, and the validation of the AR instances through Part B. Inter-rater agreement, computed on the independent codings before the consensus step, was near-perfect for presence (Po = 0.993, Pe = 0.980, Cohen's $\kappa = 0.664$; the low κ reflects the extreme prevalence of present, as only one instance was coded differently) and moderate for coverage (Po = 0.720, Pe = 0.488, $\kappa = 0.453$): 108 of 150 instances matched, and 32 of the 42 disagreements were between the adjacent categories *fully_covered* and *partially_covered*. The consensus coding produced 107 *fully_covered*, 26 *partially_covered*, 16 *missing* and 1 *wrong_mapping*. Part A yielded 148 present and 2 absent.

Recorded in Part A of the instrument, the deterministic mapping was measured against the consensus presence decisions. It reached a precision of 1.000, a recall of 0.892 and an F1 of 0.943 (132 true positives, 0 false positives, 16 false negatives, 2 true negatives), identically in Run 1 and Run 2. As there are no false positives, every error is a requirement the mapper did not find: the 16 false negatives are concentrated in AR6 (6), AR4 (5) and AR9 (3), with one each in AR7 and AR8.

Analysis of the AR instances, recorded in Part B, is limited to those the system actually validated. Of the 150 AR instances, the mapper identified 132 as present and these were passed to the local LLM; the remaining 18 were not, and 16 of them had been recorded as present by the evaluators. Of the 132 instances sent to the model, three AR5 calls returned no decision in either run, which leaves 129 validated instances. Assessment of the validation layer therefore rests on these 129 instances alone. Table 5 shows the confusion matrix for Run 1, with the human decision in rows and the system decision in columns. Measured end to end, with the 21 unvalidated instances counted as failures, agreement with the human evaluation is 36.7% in Run 1 and 34.0% in Run 2.

Table 5. Confusion matrix of system decision against human evaluation (Run 1, 129 validated instances)

Human	missing	partially	fully	Σ
missing	3	9	2	14
partially	4	17	1	22
fully	20	38	35	93
Σ	27	64	38	129

The system and the evaluators agree on 55 of the 129 decisions (42.6%, Cohen's $\kappa = 0.157$). Almost all errors lie entirely below the diagonal: the system chose a lower category in 62 cases and a higher one in only 12, which makes it on average 0.53 of a category stricter. Most disagreement is minor – 107 of the 129 decisions (82.9%) stay within one category of the human decision. This strictness is visible in the totals: the evaluators used *fully_covered* 93 times and the system only 38 times, lowering 20 of these instances as far as *missing*. Its positive decisions, however, are reliable. Of its 38 *fully_covered* decisions the evaluators confirmed 35, a precision of 0.921 against a recall of 0.376 and an F1 of 0.534. Run 2 gives the same picture: 51 decisions on the diagonal (39.5%, $\kappa = 0.134$), 66 stricter against 12 more lenient, and a precision of 0.912, a recall of 0.333 and an F1 of 0.488 for *fully_covered*.

Agreement with the human evaluation is unevenly distributed across AR nodes, as shown in Table 6.

Table 6. Agreement with human evaluation per AR node (Run 1)

AR Node	Validated Instances	Agreement	Stricter : More Lenient
AR4	19	53%	8 : 1
AR5	22	45%	10 : 2
AR6	19	21%	11 : 4
AR7	24	21%	16 : 3
AR8	23	74%	4 : 2
AR9	22	41%	13 : 0

It ranges from 74% on AR8, whose criterion is enumerable, down to 21% on AR6 and AR7, and the direction of divergence is consistent throughout: on every node the system is stricter more often than it is more lenient, most clearly on AR9, where all 13 disagreements go in the direction of strictness.

5 Discussion and Limitations

The deterministic mapper produced consistent and rule-governed AR node identification across all 25 STPs and both repeated runs. In this experiment, the behaviour of the mappers is deterministic, since mismatched instances appeared in the same position across runs. Incomplete coverage reflects the quality of the input document, e.g. short chapters, heading-only evidence, and implicitly stated content, rather than changes due to execution. These observations correspond affirmatively to RQ1.

In the context of answering RQ2, the local LLM validation layer demonstrated cross-run stability when the same evidence-based payload was submitted under identical conditions, whit decision agreement reached 94.57% and evidence sufficiency agreement 98.45% on 129 comparable pairs. Under dual validation, AR5 showed comparable stability, indicating that the additional alignment layer does not introduce systematic instability. Therefore, the cross-run consistency, including the stability of the AR5 decision under dual validation, is supported for RQ2.

Research question RQ3 is only partially answered. The evidence-based approach produced consistent and auditable distinctions between *fully_covered*, *partially_covered*, and *missing* across both runs, with all decision changes occurring between adjacent categories. The categories *wrong_mapping* and *uncertain* are operationally defined but were not triggered in the final test set. Their behaviour requires targeted evaluation on deliberately ambiguous inputs. The three core coverage categories are distinguished consistently and auditable, although the comparison with the human evaluation in Section 4.4 shows that the sufficiency threshold applied to them is miscalibrated relative to human judgement.

The answer to the last research question, RQ4, shows that the divergence of the system from the human decisions is not random but systematic. It is systematic in direction. Of the 74 disagreements in Run 1, the system assigned a lower category than the evaluators in 62 cases and a higher one in only 12; across all 129 decisions this makes the system on average 0.53 of a category stricter. Run 2 gives the same ratio, 66 to 12.

It is systematic in kind as well. When the system declares a requirement fully covered it is almost always right: of its 38 such decisions the evaluators confirmed 35 (precision 0.921). Yet of the 93 instances the evaluators considered fully covered, the system recognised only 35 (recall 0.376). The system is therefore not unreliable when it affirms, but too strict when it judges. It does not err coarsely either, as 107 of the 129 decisions stay within one category of the human decision.

The divergence is not evenly spread across requirements. Agreement is highest on AR8 (74%) and lowest on AR6 and AR7 (21%), with AR4 at 53%, AR5 at 45% and AR9 at 41%. On AR9 all 13 disagreements go in the direction of strictness and none in the direction of leniency.

Two findings show where the cause does not lie. First, in 37 of the 62 stricter disagreements the system itself recorded the evidence as sufficient and lowered the category nonetheless, so it was not short of evidence. Second, the system agrees with itself across the two runs almost perfectly ($\kappa = 0.912$) while agreeing with the evaluators only weakly ($\kappa = 0.157$), so these are not erratic lapses but a consistent shift.

Both findings point to the same conclusion: the problem lies in the threshold at which the system declares a requirement covered, not in the reliability of its inference. One possible source of the difference is that the system judges each AR node solely on the evidence segments assigned to it, whereas the evaluator reads the paper as a whole. This cannot be confirmed on the basis of the present data and requires further research.

The evaluation was conducted on 25 STPs, sufficient for controlled evaluation but not for broad generalization, and the results apply to a single local LLM configuration and prompt version. Because the comparison with the human evaluation covers the 129 validated instances and the mapper did not identify 16 requirements that the evaluators recorded as present, the quality and the structure of the input document directly affect what reaches the validation layer. Moreover, the human reference itself has limits: it was coded by the authors and settled by consensus rather than by an independent adjudicator, so it is not fully independent of the coders, and the pre-consensus coverage agreement ($\kappa = 0.453$) marks the ceiling against which the system's $\kappa = 0.157$ should be read rather than against perfect agreement. Between runs, the few decision changes are concentrated in AR7 (three) and AR8 (two). Mapping weights and acceptance thresholds were set by hand on the development and calibration subsets, and it is not known how well they transfer to other assignments or courses. The categories wrong_mapping and uncertain were not triggered by the system in the final test set, although the human evaluation produced one *wrong_mapping* decision, so their behaviour remains empirically untested. Each of these limitations calls for additional research.

6 Conclusion

This paper presented Graph2LLM, an experimental system for evaluating STP that combines deterministic graph-based mapping, evidence grounding, and a locally implemented LLM validation layer. Each validation decision is traceable to specific evidence, confidence scores, and run metadata, enabling post-hoc audit of the full evaluation process.

Experimental results demonstrate that the deterministic mapper produces stable and reproducible output, and that local LLM validation achieves cross-run stability, with variability concentrated in the most interpretive and structurally complex node. This

reproducibility was obtained with an intentionally unstable, uncensored local LLM, which indicates that cross-run stability stems from the architecture as a whole rather than from the model alone, although isolating the contribution of each component requires further research.

The results confirm that the approach distinguishes fully covered, partially covered and missing requirements reproducibly and auditably. Reproducibility and auditability are necessary but not sufficient conditions for educational assessment. A system can be highly stable and still be miscalibrated with respect to human judgement, as the comparison with the human evaluation shows. Responsibility for grading, and for its consequences, remains with the teacher and the institution; the role of the system is to support that work, not to replace it.

Future work follows from the comparison with the human evaluation. The sufficiency threshold of the validation layer will be recalibrated against human-coded examples from the development subset, since the divergence reported here is systematic rather than random. An ablation study, running the same model without the evidence payload and without the JSON schema, will show how much of the observed stability comes from the graph layer itself. The mapper will be extended with a semantic fallback for requirements that no rule maps, recorded as a probabilistic decision, so that determinism, the absence of false positives and auditability are preserved on the rule-based path. Beyond this, the same evidence-based payloads will be validated across several locally implemented models, and LLM-assisted construction of the semantic assignment layer will be explored.

References

Alghizzi, T. M., & Alshahrani, T. M. (2024). Effects of grading rubrics on EFL learners' writing in an EMI setting. *Heliyon*, *10*(18), e36394. https://doi.org/10.1016/j.heliyon.2024.e36394

Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & Von Davier, A. A. (2019). The Expanded Evidence-Centered Design (e-ECD) for Learning and Assessment Systems: A Framework for Incorporating Learning Goals and Processes Within Assessment Design. *Frontiers in Psychology*, *10*, 853. https://doi.org/10.3389/fpsyg.2019.00853

Atkinson, J., & Palma, D. (2025). An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports*, *15*(1), 14551. https://doi.org/10.1038/s41598-025-87862-3

Chiang, C.-H., Chen, W.-C., Kuan, C.-Y., Yang, C., & Lee, H. (2024). Large Language Model as an Assignment Evaluator: Insights, Feedback, and Challenges in a 1000+ Student Course. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2489–2513. https://doi.org/10.18653/v1/2024.emnlp-main.146

Directorate-General for Employment, Social Affairs and Inclusion. (2025). *DigComp 3.0 framework data supplement* [Dataset]. European Commission, Joint Research Centre. https://doi.org/10.2905/JRC.FR75K8R

Divjak, B., Svetec, B., & Horvat, D. (2024). How can valid and reliable automatic formative assessment predict the acquisition of learning outcomes? *Journal of Computer Assisted Learning*, *40*(6), 2616–2632. https://doi.org/10.1111/jcal.12953

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2024). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. (arXiv preprint arXiv:2404.16130). https://doi.org/10.48550/ARXIV.2404.16130

Elmassry, A. M., Zaki, N., Alsheikh, N., & Mediani, M. (2025). A Systematic Review of Pretrained Models in Automated Essay Scoring. *IEEE Access*, *13*, 121902–121917. https://doi.org/10.1109/ACCESS.2025.3584784

García-Varela, F., Nussbaum, M., Mendoza, M., Martínez-Troncoso, C., & Bekerman, Z. (2025). ChatGPT as a Stable and Fair Tool for Automated Essay Scoring. *Education Sciences*, *15*(8), 946. https://doi.org/10.3390/educsci15080946

Hashemi, H., Eisner, J., Rosset, C., Van Durme, B., & Kedzie, C. (2024). LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13806–13834. https://doi.org/10.18653/v1/2024.acl-long.745

Huang, Y., Wilson, J., Ormerod, C., & Beiting Parrish, M. (2025). Evaluating LLM-Based Automated Essay Scoring: Accuracy, Fairness, and Validity. *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, 71–83.

Ishida, T., Liu, T., Wang, H., & Cheung, W. K. (2024). *Large Language Models as Partners in Student Essay Evaluation*. (arXiv preprint 2405.18632). https://doi.org/10.48550/ARXIV.2405.18632

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(2), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843

Kim, Y., Hardy, M., Tey, J., Thille, C., & Piech, C. (2025). *Interpretability from the Ground Up: Stakeholder-Centric Design of Automated Scoring*

in Educational Assessments (Version 3). (arXiv preprint arXiv:2511.17069). https://doi.org/10.48550/ARXIV.2511.17069

Kostic, M., Witschel, H. F., Hinkelmann, K., & Spahic-Bogdanovic, M. (2024). LLMs in Automated Essay Evaluation: A Case Study. *Proceedings of the AAAI Symposium Series*, *3*(1), 143–147. https://doi.org/10.1609/aaaiss.v3i1.31193

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

Li, Y., Shen, X., Miao, Y., Yao, X., Ding, X., Krishnan, R., & Padman, R. (2026). *Beyond Single-Turn: A Survey on Multi-Turn Interactions with Large Language Models*. (arXiv preprint arXiv:2504.04717). https://doi.org/10.48550/arXiv.2504.04717

Li, Y., Li, Z., Wang, P., Li, J., Sun, X., Cheng, H., & Yu, J. Xu. (2024). A Survey of Graph Meets Large Language Model: Progress and Future Directions. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 8123–8131. https://doi.org/10.24963/ijcai.2024/898

Liu, T., Ye, L., & Yan, W. (2026). A framework for evaluation of large language models in essay assessment: Reliability, alignment, and causal reasoning. *Computers and Education: Artificial Intelligence*, *10*, 100565. https://doi.org/10.1016/j.caeai.2026.100565

Ma, C., Chen, Y., Wu, T., Khan, A., & Wang, H. (2025). Large Language Models Meet Knowledge Graphs for Question Answering: Synthesis and Opportunities. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 24589–24608. https://doi.org/10.18653/v1/2025.emnlp-main.1249

Mughal, N., Imran, A. S., Daudpota, S. M., Kastrati, Z., & Noor, W. (2026). Exploring potential of large language models for automated essay scoring in education. *Discover Artificial Intelligence*, *6*(1), 166. https://doi.org/10.1007/s44163-026-01002-y

Narreddy, C., Joordens, S., & Prompiengchai, S. (2025). Harnessing Large Language Models for Scalable and Effective Formative Assessment in Higher Education: A Review. *Trends in Higher Education*, *4*(4), 65. https://doi.org/10.3390/higheredu4040065

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, *6*, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2023). *Unifying Large Language Models and Knowledge Graphs: A Roadmap*. (arXiv preprint arXiv:2306.08302) https://doi.org/10.48550/ARXIV.2306.08302

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., & Henderson, P. (2024). Fine-tuning aligned language models compromises safety, even when users do not intend to! In *Proceedings of the Twelfth International Conference on Learning Representations* (ICLR 2024). 30988–31043. https://proceedings.iclr.cc/paper_files/paper/2024/file/83b7da3ed13f06c13ce82235c8eedf35-Paper-Conference.pdf

Qu, K., Li, K. C., Wong, B. T. M., Wu, M. M. F., & Liu, M. (2024). A Survey of Knowledge Graph Approaches and Applications in Education. *Electronics*, *13*(13), 2537. https://doi.org/10.3390/electronics13132537

Song, Y., Zhu, Q., Wang, H., & Zheng, Q. (2024). Automated Essay Scoring and Revising Based on Open-Source Large Language Models. *IEEE Transactions on Learning Technologies*, *17*, 1880–1890. https://doi.org/10.1109/TLT.2024.3396873

Tannous, M. E., Ramadan, W. H., & Rajab, M. A. (2023). TSHD: Topic Segmentation Based on Headings Detection (Case Study: Resumes). *Advances in Human-Computer Interaction*, *2023*, 1–12. https://doi.org/10.1155/2023/6044007

University of Zadar. (2026). Policy on the Use of Artificial Intelligence at the University of Zadar. https://www.unizd.hr/Portals/0/doc/doc_pdf_dokumenti/pravilnici/politika_koristenja_umjetne_inteligencije_na_sveucilistu_u_zadru_20260225.pdf?ver=iaHevmwh7XPEjZAGSF0lOw%3D%3D

Vangelova, A., & Gancheva, V. (2026). AI-Based Automated Scoring Layer Using Large Language Models and Semantic Analysis. *Applied Sciences*, *16*(7), 3537. https://doi.org/10.3390/app16073537

Wang, Y., Ding, Z., Wu, X., Sun, S., Liu, N., & Zhai, X. (2026). AutoSCORE: Enhancing Automated Scoring with Multi-Agent Large Language Models via Structured Component Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(48), 40898–40906. https://doi.org/10.1609/aaai.v40i48.42123

Xu, W., Kassim, M. S. S., Hoo, W. L., Yang, W., & Xu, T. (2026). Explainable AI for education: Enhancing essay scoring via rubric-aligned chain-of-thought prompting. *International Journal of Modern Physics C*, *37*(06), 2542013. https://doi.org/10.1142/S0129183125420136

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of*

Educational Technology, *56*(1), 150–166. https://doi.org/10.1111/bjet.13494

Young, R. J. (2026). Comparative analysis of LLM abliteration methods: A cross-architecture evaluation (arXiv preprint arXiv:2512.13655). https://doi.org/10.48550/arXiv.2512.13655

Zapata-Rivera, D., Forsyth, C., Graf, E., & Jiang, Y. (2024). *Designing and Evaluating Evidence-Centered Design based Conversations for Assessment with LLMs*. https://par.nsf.gov/biblio/10591735-designing-evaluating-evidence-centered-design-based-conversations-assessment-llms

Zesch, T., Hanses, M., Seidel, N., Aggarwal, P., Veiel, D., & de Witt, C. (2024). *FernUni LLM Experimental Infrastructure (FLEXI)—Enabling Experimentation and Innovation in Higher Education Through Access to Open Large Language Models* (Version 1). (arXiv preprint arXiv:2407.13013). https://doi.org/10.48550/ARXIV.2407.13013

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena* (Version 4). (arXiv preprint arXiv:2306.05685). https://doi.org/10.48550/ARXIV.2306.05685