Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Curriculum-Aligned Analysis of Assessment Feedback: A Retrieval-Augmented Large Language Model Approach for Revealing Fine-Grained Student Error Patterns in Higher Education

Linxin Hua

linxin.hua1@monash.edu

Monash University

Jia Xu

Deakin University

Lirui Guo

Monash University

Nan Zheng

Monash University

Qingtan Shen

University of Melbourne

Dilang Tan

University of Melbourne

Ye Lu

Monash University

Research Article

Keywords: Assessment feedback, Large language models, Retrieval-augmented generation, Curriculum-aligned analysis, Higher education

Posted Date: June 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9561394/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Assessment feedback in higher education provides fine-grained evidence about students' specific errors but remains underutilized as a data source for teaching improvement. This study proposes a method based on a large language model (LLM) that transforms unstructured feedback into curriculum-aligned error analysis through three steps: data-driven generation of an error taxonomy from feedback content, mapping each piece of feedback to curriculum content via retrieval-augmented generation, and aggregating annotations into error patterns with quantified cohort coverage. The method was applied to 6072 pieces of feedback across four cohorts (2022–2025) of an undergraduate engineering course. The resulting annotations achieved 89.1% agreement with human judgment, substantially higher than the 74.7% obtained when an expert-designed taxonomy was applied to the same data. In a focus group, members of the teaching team confirmed the output as consistent with their marking experience, used the cross-year trends to reflect on the impact of past teaching adjustments, and proposed targeted curriculum revisions; they also cautioned that error rates should not be repurposed as a measure of teaching quality. The findings reframe assessment feedback as a reflective resource for teaching teams, one that reveals the impact of past instructional changes and informs the reallocation of teaching effort.

1. Introduction

Teachers typically draw on assessment results and rubric-based statistics to monitor student understanding of learning materials (Schildkamp, 2019). These resources can indicate that students are struggling but rarely reveal which specific concepts or procedures students misunderstand, or why those difficulties arise and how instruction might be adjusted to address them. For example, rubric-based statistics, e.g., the percentage of students scoring below a threshold on a given criterion can capture student performance across a cohort (Panadero et al., 2023), yet offer limited insight into the nature or origins of their difficulties.

Assessment feedback, by contrast, describes student errors at the level of specific concepts or procedures. A comment such as "missing justification for the chosen design speed", for example, simultaneously identifies a curriculum topic (design speed), the type of error (a missing justification), and the link between them. Feedback may therefore contain the information to reveal which curriculum content is associated with specific errors and how those associations arise. Thus, the analysis of feedback may enable teachers to make targeted improvements to their instruction. However, feedback is predominantly viewed as a tool for supporting and assessing student learning rather than as a data source for informing teaching practices (Baartman & Quinlan, 2024; Boud & Molloy, 2013).

Realizing this potential is far from straightforward. Feedback is typically unstructured text that cannot be directly counted, averaged, or compared, and requires interpretation before any meaningful aggregation is possible (Shermis & Wilson, 2024). In courses with hundreds of students, manually extracting cohort-level patterns is infeasible, and although teachers may develop intuitive impressions during marking, translating these into verifiable insights requires considerable effort (Phung & Michell, 2022; Vanlommel

et al., 2017). When marking is distributed across multiple staff, feedback further varies in writing style, level of detail, and focus (Bloxham et al., 2016; Derham et al., 2022), compounding the challenge.

Recent advances in large language models (LLMs) have enabled automated analysis of unstructured text at scale (Alizadeh et al., 2025; Carlson & Burbano, 2026). However, applying such methods to assessment feedback presents distinct challenges: feedback covers diverse topics, the same error behavior can occur across different topics, and transforming this unstructured information into curriculum-aligned insights requires not only classifying errors but also linking them to specific curriculum content. To date, no existing method combines error classification with curriculum-level mapping in feedback analysis at scale.

Beyond the design of such an approach, the reliability of its output requires careful examination. Categories developed for human interpretation may not be directly applicable for LLM-based classification, as LLMs are highly sensitive to how categories are defined and even minor wording changes can shift classification outcomes (Dunivin, 2025; Halterman & Keith, 2026). Whether an LLM-based approach can produce outputs that align with human judgment, and under what conditions discrepancies arise, has yet to be established.

Beyond classification reliability, a further question is whether the resulting analysis is practically useful for teaching improvement. Prior research has highlighted a recurring gap between the data available to teachers and the data they need for curriculum decisions: current learning analytics tools primarily present engagement metrics and assessment scores, which indicate where students struggle but not what specific errors they make (Claassen et al., 2025; Schildkamp, 2019). Whether curriculum-aligned error patterns can provide this specificity, and how teaching teams interpret and act upon such evidence, has not been explored.

To address these gaps, this study proposes an LLM-based method that transforms unstructured assessment feedback into curriculum-aligned error analysis. The method is applied to feedback from an undergraduate engineering course. The following research questions guide this study:

1. How can unstructured assessment feedback be used for curriculum-aligned error analysis?
2. To what extent do the method's outputs align with human judgment, and where do discrepancies arise?
3. How does the teaching team perceive the curriculum-aligned error patterns, and in what ways do these patterns inform curriculum improvement decisions?

Through this study, we aim to advance the use of assessment feedback as a data source for teaching improvement. We develop and validate an LLM-based method for curriculum-aligned feedback analysis, applying it to 6072 pieces of feedback across four cohorts (2022–2025) of an undergraduate engineering course and reviewing the output with members of the teaching team in a focus group. We examine the alignment between AI-derived classification and educators' professional judgment, and how teaching teams engage with such evidence in interpreting and acting upon cohort-level student error

patterns. The study offers empirical findings and practical considerations for instructors, course coordinators, curriculum designers, and learning analytics researchers seeking to leverage qualitative educational data to inform teaching decisions in higher education.

2. Literature Review

2.1. Assessment Feedback as a Source of Teaching Insight

Feedback is recognized as one of the most influential factors in student learning (Black & Wiliam, 1998; Wisniewski et al., 2020), and research has accordingly focused on how to make it more effective for learners. Studies have examined the skills and knowledge that students need to act on feedback (Carless & Boud, 2018) and that teachers need to design effective feedback processes (Boud & Molloy, 2013; Istencioglu et al., 2026). Baartman and Quinlan (2024) argued that feedback should be coordinated across all assessments in a course so that each round of feedback supports ongoing student development. Across these studies, feedback is consistently conceptualized as a communication channel directed toward students, and research attention remains on how it is designed, delivered, received, and acted upon by learners.

Recent efforts to analyze feedback at scale have maintained this student-facing orientation. Systematic literature reviews of learning analytics for feedback in higher education found that analytics has primarily been employed to enhance the feedback itself, focusing on practices such as monitoring student engagement with feedback, identifying at-risk learners, and personalizing feedback delivery to individual students (Banihashem et al., 2022; Banihashem et al., 2025). Feedback has also been analyzed along quality dimensions such as specificity and constructiveness (Cavalcanti et al., 2020; Osakwe et al., 2022), examined for its linguistic features (Derham et al., 2022), and classified according to what aspect of student work it addresses (Aldino et al., 2025; Lin et al., 2023). These studies all evaluate the quality of feedback rather than mine its content for information about what errors students made.

The information that feedback contains is qualitatively different from what other assessment data sources can provide. For example, assessment scores, rubric ratings, and learning analytics dashboards capture student performance at the level of overall criteria or engagement indicators (Viberg et al., 2018), but a grade or score constitutes a compressed summary that describes how well students performed without explaining what specific mistakes were made or why (Sadler, 1989). Feedback written by markers, by contrast, identifies the specific strengths and weaknesses in students' work, for instance, that a student applied an incorrect procedure, omitted a required justification, or misinterpreted a key concept (Nicol & Macfarlane-Dick, 2006).

Despite this information richness, feedback remains largely unexplored as a data source for teaching improvement. Its potential is constrained by the unstructured nature of feedback text, which cannot be directly quantified, compared, or summarized without interpretation (Shermis & Wilson, 2024), and in courses with multiple markers, variation in writing style, level of detail, and focus introduces further

inconsistency (Bloxham et al., 2016). Assessment feedback therefore remains an underutilized data source whose information granularity has not been systematically leveraged for curriculum-level error analysis.

2.2. Analyzing Assessment Feedback at Scale

Given the practical constraints identified above, extracting systematic insights from feedback requires methods that can process unstructured text at scale. Manual approaches such as content analysis and qualitative coding can structure feedback text into interpretable categories (Krippendorff, 2019), but these methods are labor-intensive and difficult to scale beyond small samples. Although teachers may develop intuitive impressions of common errors during marking, translating these impressions into systematic and verifiable insights requires considerable additional effort (Phung & Michell, 2022; Vanlommel et al., 2017).

Advances in natural language processing (NLP) and LLMs have enabled automated text analysis at scale, with recent studies showing that LLMs can match or exceed human performance on text annotation (Alizadeh et al., 2025; Carlson & Burbano, 2026). In educational contexts, earlier NLP approaches addressed automated essay scoring (Shermis & Wilson, 2024) and classification of student writing (McNamara et al., 2014), and LLMs have since been applied to feedback generation, tutoring, and student response analysis (Kasneci et al., 2023; Yan et al., 2024). For student errors specifically, Kökver et al. (2025) trained supervised classifiers to detect misconceptions about the greenhouse effect, achieving 88.9% accuracy across three pre-specified classes.

However, applying LLM-based text analysis to assessment feedback presents distinct challenges. Feedback spans diverse curriculum topics and error categories, and the same error behavior (e.g., missing justification) can appear across different topics. Transforming this unstructured information into actionable teaching insights therefore requires not only classifying errors but also linking them to specific curriculum content. To date, no existing method addresses this combined challenge of error classification and curriculum-level mapping in assessment feedback analysis.

2.3. Error Classification and Curriculum Alignment

Effective error classification depends on the quality of the categories employed. Research on student errors has traditionally relied on expert-designed taxonomies (Shimizu & Kang, 2025), which are typically qualitative, small-scale, and domain-specific, making them difficult to apply across courses or to scale. An alternative is to derive categories inductively from the data, reducing dependence on prior expertise. LLMs can assist in such data-driven taxonomy construction, producing codebooks comparable to expert-designed ones through human-LLM collaboration (Dai et al., 2023; Than et al., 2025; Wan et al., 2024), though LLM-generated codes can remain surface-level and may require human oversight (Parfenova et al., 2025; Vikan et al., 2026).

However, classifying feedback text into general error categories does not indicate where in the curriculum the classified errors originate (Shulman, 1986). The principle of constructive alignment holds

that learning outcomes, teaching activities, and assessment should be coherently linked (Biggs et al., 2022). Since assessment feedback is a direct product of this alignment, reflecting how student performance relates to intended learning outcomes, the errors captured in feedback should similarly be traceable back to the specific curriculum content from which those outcomes derive. Emerging research has explored the use of LLMs to support constructive alignment at the course-design level, for example, generating aligned learning outcomes and assessment tasks (Pereira et al., 2024), but mapping observed student errors back to the curriculum remains unaddressed. Curriculum mapping practices have examined how learning outcomes align across courses and programs (Arafeh, 2016), yet these approaches operate at the program level rather than tracing individual errors to specific taught content. Existing NLP applications to student responses, such as Mello et al.'s (2022) clustering of open-ended answers by textual similarity, similarly do not connect resulting patterns to specific taught content. A mechanism for establishing traceable links between classified errors and curriculum content is therefore needed.

Retrieval-augmented generation (RAG) offers an approach for establishing such curriculum-level connections by grounding LLM outputs in source material retrieved from a knowledge base (Fan et al., 2024; Lewis et al., 2020). Recent educational applications have demonstrated that course materials can serve as an effective knowledge base: Fung et al. (2024) and Jacobs and Jaschke (2024) used RAG to generate feedback aligned with teaching materials, while Li et al. (2024) matched test questions to curriculum goals using textbook content. A systematic review confirms that semantic retrieval from course materials can establish traceable links between text and curriculum content (Li et al., 2025). However, RAG has primarily been employed for student-facing functions rather than teaching-facing analysis.

Furthermore, even when individual feedback can be classified and linked to curriculum content, the resulting annotations must be aggregated into patterns that instructors can act upon. Feedback describes errors in natural language where different markers may use different words for the same issue and similar phrasing may refer to different problems. Before such information can be quantified and organized by curriculum structure, semantically similar descriptions must be identified and grouped. However, to the best of our knowledge, there is no existing method combines error classification with curriculum-level mapping and consolidates such mapping into quantified error patterns for use in teaching practice.

2.4. Reliability of LLM-Based Classification

Beyond the design of an error classification and curriculum-mapping approach, the reliability of its output requires careful examination. Research has demonstrated that LLMs are highly sensitive to the wording of category definitions. Dunivin (2025) found that codebooks designed for human coders must be substantially rewritten for LLM use, because they contain implicit knowledge that LLMs cannot access, and even minor wording changes can shift classification outcomes significantly. Halterman and Keith (2026) similarly reported that manually written category definitions do not reliably improve LLM classification and can even introduce confusion.

These findings raise questions about whether categories designed through expert judgment are directly suitable for LLM-based classification. Data-driven taxonomies, by contrast, are generated from the same textual patterns that the LLM will subsequently classify, which may produce definitions more compatible with LLM processing. However, whether this compatibility translates into higher classification agreement with human judgment has not been empirically established.

Evaluating the reliability of LLM-based classification also requires appropriate validation methods. Inter-rater reliability metrics such as Cohen's κ (Cohen, 1960) have been adopted in recent LLM annotation studies to quantify agreement between LLM and human annotations (Alizadeh et al., 2025; Carlson & Burbano, 2026). However, validation practices for LLM-based classification in educational research remain limited, and whether expert-designed or data-driven taxonomies produce more reliable results when applied to assessment feedback has not been established.

2.5. Data-Informed Teaching and Curriculum Improvement

Whether quantified analytics outputs translate into curriculum-improvement decisions depends on how teaching teams interpret them in their own teaching context. Research on data-informed teaching practice has consistently emphasized that effective instructional decisions draw on a combination of structured evidence, contextual knowledge of the cohort, and the teacher's own experience of how teaching unfolded (Mandinach & Schildkamp, 2021; Schildkamp, 2019). Ho (2022) observed that teachers routinely rely on tacit qualitative observations to make instructional choices, but these observations are rarely recorded or analyzed in a structured form. Existing teacher-facing analytics tools provide such a record only at the level of aggregated grades, engagement, or rubric scores organized by module (Echeverria et al., 2018; Milesi & Martinez-Maldonado, 2024), and how teaching teams integrate these outputs into curriculum-revision decisions has been studied less than the design of the tools themselves.

Among current applications of automated qualitative analysis to teaching content, the analysis of student course evaluations is one of the main lines of work. Gottipati et al. (2018) applied text analytics to extract course improvement suggestions from student survey responses, and Cai et al. (2025) employed a multi-agent LLM framework to generate institutional reports from evaluation data. These studies establish that automated qualitative analysis can produce evidence intended for teaching teams, but course evaluations capture student perceptions of the learning experience rather than the specific errors students make in their assessed work. The interpretive task a teaching team faces when reading content-specific, curriculum-aligned error evidence drawn from their own marking may differ substantially.

How teaching teams perceive curriculum-aligned error patterns derived from feedback, and how such patterns inform their reasoning about curriculum changes, therefore remains an open empirical question. Addressing it requires examining how a teaching team familiar with the course content engages with such evidence in an authentic decision-making context.

3. Method

3.1. Overview

3.1.1. Method Workflow

To address the gaps, this study aims to transform unstructured assessment feedback into curriculum-aligned error analysis through three sequential steps. Step 1 is error taxonomy generation, which derives error categories from the feedback data. Step 2 is error-to-curriculum mapping, which annotates each piece of feedback with an error category and a link to specific curriculum content. Step 3 is pattern aggregation and coverage analysis, which consolidates individual annotations into curriculum-aligned error patterns, quantifies their coverage across the cohort, and organizes the results by teaching sequence. Figure 1 illustrates the overall workflow.

3.1.2. Case Context and Data

The method was applied to a third-year undergraduate engineering course in which students work in groups of five to six to complete a semester-long design project. The project is assessed through two written report submissions at different design stages (Stage A and Stage B), each covering different aspects of the design process but requiring similar skills in analysis, calculation, and justification. Each semester, approximately 200–300 students enroll in the course, forming 40–60 groups.

Marking was carried out by five to seven teaching assistants with relevant academic backgrounds, following rubrics with predefined criteria (12 for Stage A, 9 for Stage B). Before each marking period, the teaching team held calibration sessions using sample submissions, and feedback was required to be detailed, actionable, and specific in identifying what students did wrong.

Table 1 summarizes the dataset, which spans four cohorts from 2022 to 2025 (6072 pieces of feedback across 322 student groups). The same three-step workflow was applied to all cohorts. The 2022 dataset (1738 pieces across 49 groups) is used as the running example throughout Section 3 to illustrate intermediate outputs; cross-year results across all cohorts are presented in Section 4.

Table 1
Dataset overview

Year	Stage	Groups	Feedback / stage	Feedback / year
2022	A	49	1041	1738
2022	B	49	697	
2023	A	37	710	1283
2023	B	40	573	
2024	A	39	1457	2123
2024	B	34	666	
2025	A	37	642	928
2025	B	37	286	
Total		322	6072	6072

3.2. Error Taxonomy Generation

Feedback was first organized into categories that capture the types of errors students make. Rather than relying on predefined categories, we adopted a data-driven approach that derives them from the feedback itself, removing the need for prior familiarity with the course content.

The taxonomy generation process shown in Fig. 2 was applied to all 2022 feedback data (Stages A and B, 1738 pieces of feedback) using gpt-5.4-nano-2026-03-17. The process comprises four sub-steps: three LLM sub-steps (Steps 1.1, 1.2, and 1.4) separated by one human review (Step 1.3).

In Step 1.1, each piece of feedback was abstracted into a behavioral characterization describing the underlying error behavior independent of the subject matter. For example, "The calculation procedure is not explained" was characterized as "Missing procedural transparency". Removing curriculum-specific content allowed behaviors to be compared and grouped across topics. This sub-step produced 258 unique behavioral characterizations from the 1738 pieces of feedback.

In Step 1.2, the LLM grouped these 258 characterizations into merged categories based on semantic similarity, producing eight candidate categories.

In Step 1.3, a human reviewer (an academic staff member with no prior involvement in the course, chosen to test whether the method works without domain expertise) examined the candidate groupings, merged three categories addressing absence of required elements into a single **Incomplete / Missing Content & Justification** category, and reassigned 26 characterizations, producing five final categories.

In Step 1.4, the LLM generated, for each final category, a definition summarizing its behavioral characterizations and a boundary contrasting it against the other categories.

Table 2 presents the resulting taxonomy, comprising four error categories and one **Positive** category for feedback confirming correct work. This taxonomy is embedded in prompts for the subsequent mapping step. Because the assessment structure, rubric criteria, and required skills remain consistent across years, the taxonomy was applied to all subsequent cohorts (2023–2025) in the cross-year analysis.

Table 2
Error taxonomy generated from 2022 feedback data

Category	Definition	Boundary
Incomplete / Missing Content & Justification	Essential content, data, or justification is absent due to cognitive or procedural omissions, resulting in incomplete analysis, design, or documentation.	The category is defined by the explicit absence of required elements rather than flaws in what was produced. It is distinguished from Procedural / Process Flaws, which concerns execution quality rather than absence, and excludes cases where content is present but incorrectly reasoned (Reasoning & Interpretation Errors) or poorly presented (Presentation / Communication).
Procedural / Process Flaws	Errors in how work is executed, including skipped steps, misapplied methods, incorrect rounding, and failure to follow formatting or documentation requirements.	The category captures breakdowns in execution when the required elements are present but handled incorrectly. It excludes purely interpretive mistakes (Reasoning & Interpretation Errors), issues in how results are conveyed (Presentation / Communication), and the wholesale absence of required content (Incomplete / Missing Content & Justification).
Reasoning & Interpretation Errors	Cognitive mistakes in understanding concepts, criteria, or data, manifesting as misinterpretations, faulty inferences, or unjustified assumptions that produce incorrect conclusions.	The category captures flawed interpretation when content and procedure are otherwise sound. It excludes procedural slips (Procedural / Process Flaws), delivery issues (Presentation / Communication), and problems arising from missing information rather than misinterpretation (Incomplete / Missing Content & Justification).
Presentation / Communication	Errors in how information is presented or communicated, including formatting, visual design, and clarity issues, where the underlying content exists but is conveyed ineffectively.	The category focuses on how content is delivered rather than what it contains or how it was reasoned. It excludes errors in content logic or interpretation (Reasoning & Interpretation Errors), missing content (Incomplete / Missing Content & Justification), and breakdowns in the execution process itself (Procedural / Process Flaws).
Positive	Feedback expressing approval or confirming correct work, with no error identified.	The category includes explicit praise, confirmation of correct results, and affirmative feedback that conveys satisfaction without describing problems. It excludes any content identifying errors, omissions, or need for improvement.

3.3. Error-to-Curriculum Mapping

With an error taxonomy established, Step 2 links each error to the specific curriculum content it relates to. This is implemented through two components: a curriculum knowledge base built from the course's teaching materials, and a RAG process that maps each piece of feedback to the most relevant knowledge unit.

3.3.1. Knowledge Base Construction

The knowledge base was constructed from the course's lecture slides, organized into knowledge units (self-contained segments each addressing a single topic; Nafa et al., 2016). Slides were processed page-by-page with a multimodal LLM (gpt-5.4-mini-2026-03-17), which identified topic boundaries from the visual layout and textual content. A teaching assistant with experience in the course reviewed every candidate knowledge unit and adjusted boundaries where a topic spanned multiple pages or a single page contained more than one topic. The resulting knowledge base contains 374 knowledge units, each linked to its source slide and page range, and embedded using OpenAI's text-embedding-3-large model.

3.3.2. Mapping Process

The mapping process uses RAG to link each piece of feedback to relevant curriculum content through two sub-steps, shown in Fig. 3. In Step 2.1, the feedback is embedded using the same model as the knowledge units. Cosine similarity is then computed between the feedback embedding and all 374 knowledge units in the knowledge base, and the five units with the highest similarity scores are retrieved as candidates. In Step 2.2, the LLM (gpt-5.4-nano-2026-03-17) receives the feedback, the five candidate units, and the error taxonomy (as shown in Table 2), and produces an annotation that links the feedback to the most relevant knowledge unit, assigns an error category, and provides a reasoning explanation. As each piece of feedback is processed independently, the mapping scales linearly and can be parallelized.

The reasoning in each annotation explains why the feedback reflects the assigned category in the context of the linked knowledge unit. In Fig. 3, for example, the feedback "Modest and implicit discussion of curve coordination" is annotated as **Incomplete / Missing Content & Justification** because the linked knowledge unit defines the concept students were required to address explicitly. These reasoning descriptions serve as the basis for Step 3. Feedback classified as **Positive** was not linked to any knowledge unit. All 1738 pieces of feedback from the 2022 dataset were processed through this pipeline.

3.4. Pattern Aggregation and Coverage Analysis

The mapping step produces an annotation for each piece of feedback, but a single annotation describes only one feedback item, which is too narrow for instructors to identify course-level patterns to act upon. Step 3 consolidates these annotations into error patterns, each of which is a single textual description of one recurring student error, scoped to one teaching week and potentially spanning multiple knowledge units. Cohort coverage is then computed for each pattern through three sub-steps, shown in Fig. 4.

The 1255 error-bearing annotations were first organized into 397 (knowledge unit, error category) groups. In Step 3.1, the LLM identified semantically similar reasoning descriptions within each group and consolidated them into merged annotations, addressing the variation in how different markers describe the same error. Only groups with at least three annotations were processed, both because semantic clustering requires multiple instances and because the analysis aimed to focus on errors that affected more students. 282 groups containing 1–2 annotations each (350 reasoning descriptions in total) were therefore excluded from subsequent steps. Across the 115 processed groups, this step reduced 905 reasoning descriptions to 194 merged annotations.

In Step 3.2, the LLM consolidates related merged annotations across different knowledge units within the same teaching week into error patterns. The same type of student error often relates to multiple curriculum topics, producing separate merged annotations under different but related knowledge units during mapping, and Step 3.2 groups these dispersed annotations under one pattern description. The mapping between knowledge units and patterns is therefore many-to-many. As illustrated for Week 4 in Fig. 4, the highlighted pattern draws from three knowledge units, and unit 05-21-00 splits its annotations between two patterns. Across the 2022 dataset, this step further consolidated 194 merged annotations into 60 error patterns.

In Step 3.3, cohort coverage is computed for each error pattern as the percentage of student groups whose feedback relates to that pattern, and results are organized by teaching week. Using percentages rather than raw counts keeps results comparable across cohorts of different sizes. The output combines qualitative descriptions with their quantitative prevalence, allowing instructors to identify which weeks and topics require adjustment.

For the cross-year analysis, the full pipeline was applied independently to all eight datasets (four cohorts with two stages each) from 2022 to 2025 using the same taxonomy and knowledge base. Error patterns produced for the same stage and teaching week were then matched across cohorts: patterns describing the same underlying issue were grouped under a unified label using the semantic-similarity consolidation introduced in Step 3.2. The resulting output combines per-cohort weekly summaries with year-by-year coverage trajectories for these recurring patterns.

3.5. Evaluation Design

The method was evaluated from two complementary perspectives. Section 3.5.1 examines classification reliability (RQ2), testing whether the data-driven taxonomy produces annotations that align with human judgment on the 2022 dataset. Section 3.5.2 examines practical utility (RQ3) through a focus group with the teaching team, testing whether the cross-year analysis output (covering 2022 to 2025) is consistent with teaching experience and useful for informing curriculum decisions.

3.5.1. Classification Validation (RQ2)

To evaluate whether the data-driven taxonomy supports reliable error classification, we compared it against an independently developed expert taxonomy. The expert taxonomy was designed by the unit coordinator based on the understanding of common student errors in this course, without reviewing the feedback. It contains seven categories (Calculation, Misconception, Omission, Procedural, Careless, Positive, and Other). As with the data-driven taxonomy, each category was accompanied by a definition and an inter-category boundary; in the expert taxonomy these were authored manually by the unit coordinator rather than generated by an LLM. The two taxonomies therefore shared the same structural form (label, definition, boundary) but differed in their derivation, with the data-driven categories emerging inductively from the feedback and the expert categories specified a priori from expert knowledge. The same 1738 pieces of feedback were processed using both taxonomies through the mapping pipeline described in Section 3.3, and both sets of results underwent a two-phase validation process.

In the first phase, each annotation record was independently reviewed five times by gpt-5.4-mini-2026-03-17, a more advanced model than the one used for annotation. For each record, the reviewer model assessed whether the assigned error category and knowledge unit were appropriate for the given feedback. When all five runs agreed with the pipeline's annotation, the record was marked as confirmed and exempted from mandatory human review. Otherwise, it was flagged for human examination.

In the second phase, all flagged records and a stratified random sample of unflagged records were reviewed by an experienced teaching assistant with three years of marking experience in the course. The reviewer was shown the feedback text, the pipeline's linked knowledge unit, and the taxonomy under evaluation, and selected the most appropriate error category and judged whether the linked knowledge unit was relevant, without seeing the pipeline's category assignment or flag status. Showing the linked knowledge unit kept validation tractable while preserving independent judgment on the error category. The stratified sample drew 20 records per error category (including **Positive**), with smaller categories sampled in full. Agreement between the mapping pipeline's error category assignment and the human reviewer's selection was measured using Cohen's κ (Cohen, 1960) and percentage agreement. Agreement on knowledge-unit relevance, defined as whether the reviewer judged the knowledge unit linked by the mapping pipeline as relevant to the feedback, was measured as percentage agreement. The five-run LLM screening served to flag records for human review and did not contribute to these agreement measures.

3.5.2. Teaching Team Review (RQ3)

To examine whether the curriculum-aligned error patterns are consistent with teaching experience and useful for informing curriculum decisions, a focus group discussion was conducted with three members of the teaching team: the unit coordinator (UC) and two teaching assistants (TA1, TA2), collectively covering all four cohorts with three to four years of marking experience. None of the participants were involved in the classification validation or had prior exposure to the pipeline's output.

The procedure followed human-centred learning analytics principles, under which analytics tools are evaluated through how educators engage with them in authentic decision-making contexts (Buckingham

Shum et al., 2019; Klein et al., 2019). Drawing on Wise and Jung's (2019) situated model of instructor analytics use, the focus group probed four aspects of engagement: validity, novelty, interpretability, and actionability.

Participants were provided with the complete cross-year analysis output as a printed report containing ranked error patterns, trend charts, and weekly detail with year-by-year coverage. After individual reading, a semi-structured focus group probed whether the error patterns matched participants' marking experience, whether any patterns surprised them in scale or ranking, whether the cross-year trends aligned with known changes in teaching materials or delivery, and how they would use the analysis for curriculum improvement. The hour-long session was audio-recorded and transcribed, then analyzed using inductive thematic analysis (Braun & Clarke, 2006), with themes supported by verbatim quotes.

4. Results

This section presents the results organized by research questions. Section 4.1 addresses RQ1 by presenting the curriculum-aligned error analysis produced by the three-step method. Section 4.2 addresses RQ2 by quantifying how closely the pipeline's annotations align with human expert judgment under data-driven and expert taxonomies, and by identifying where the remaining discrepancies are concentrated. Section 4.3 addresses RQ3 by reporting how the teaching team interprets the curriculum-aligned error patterns and how these patterns inform curriculum improvement decisions.

4.1. Curriculum-Aligned Error Analysis (RQ1)

4.1.1. Error Classification

Of the 1738 pieces of feedback in the 2022 dataset, 483 were classified as **Positive**, leaving 1255 error-bearing pieces. Table 3 presents their distribution across the four error categories. **Incomplete / Missing Content & Justification** is the dominant category, accounting for approximately two-thirds of all errors. The distribution differs between the two stages: **Incomplete / Missing Content & Justification** accounts for 73.3% in Stage A but drops to 54.2% in Stage B, while **Procedural / Process Flaws** rises from 13.2% to 30.3%.

Table 3
Distribution of error-bearing feedback by error category (data-driven taxonomy)

Error Category	Stage A (n = 810)	Stage B (n = 445)	Combined (n = 1255)
Incomplete / Missing Content & Justification	594 (73.3%)	241 (54.2%)	835 (66.5%)
Procedural / Process Flaws	107 (13.2%)	135 (30.3%)	242 (19.3%)
Presentation / Communication	78 (9.6%)	32 (7.2%)	110 (8.8%)
Reasoning & Interpretation Errors	31 (3.8%)	37 (8.3%)	68 (5.4%)

4.1.2. Curriculum-Aligned Mapping

The curriculum-aligned mapping reveals where these errors concentrate within the taught content. All 1255 error-bearing pieces of feedback were linked to knowledge units, mapping to 200 of the 374 available knowledge units across 39 of the 40 lecture slide modules. Figure 5 shows the distribution across modules for each stage, organized by teaching sequence. While most modules are dominated by **Incomplete / Missing Content & Justification**, certain modules show distinct profiles. For instance, in Stage B, modules W9.2 (Design traffic Pt 2) and W10.1 (Empirical chart design method) each have over half their error-bearing feedback classified as **Procedural / Process Flaws**.

4.1.3. Aggregation and Weekly Summaries

Table 5 presents the weekly summaries produced by Step 3 for the 2022 dataset. The Stage A and Stage B coverage columns report the weekly cohort coverage, that is, the percentage of student groups whose feedback for that week mentioned at least one identified error pattern. This weekly coverage varies substantially across teaching weeks: feedback related to Weeks 1, 4, and 5 in Stage A covers over 89% of student groups, while Week 2 covers only 14.3%. The rightmost column shows the most prevalent error pattern in each week, paired with its per-pattern cohort coverage (the percentage of groups whose feedback referenced that specific pattern), pointing instructors to which topic most students struggled with that week.

Table 5
Weekly aggregation results for 2022

Week	Stage A Coverage	Stage A KUs	Stage A Patterns	Stage B Coverage	Stage B KUs	Stage B Patterns	Most Prevalent Error Pattern (Coverage)
1	91.8%	5	6	46.9%	3	7	Gaps in core economic evaluation and incremental analysis (51.0%)
2	14.3%	1	1	-	-	-	Missing justification for chosen design speed (14.3%)
3	61.2%	5	5	-	-	-	Missing core horizontal curve data and calculation methods (34.7%)
4	89.8%	8	7	-	-	-	Comprehensive curve coordination coverage and justification (67.3%)
5	98.0%	12	7	63.3%	2	3	Comprehensive cross-section data gaps (93.9%)
6	46.9%	5	6	-	-	-	Missing drawings, diagrams, and plan-view annotations for intersections (26.5%)
7	-	-	-	28.6%	3	3	Missing justification for pavement design, materials, and scenario factors (22.4%)
8	-	-	-	20.4%	1	4	Missing or unclear CBR data, mappings, and justifications (14.3%)
9	-	-	-	63.3%	4	5	DESA/HVAG/ESA data integrity and downstream effects (22.4%)
10	-	-	-	59.2%	3	6	Inconsistent thickness rounding and incomplete calculation procedure (40.8%)

Stage A/B Coverage = weekly cohort coverage, the percentage of the 49 student groups whose feedback for that week mentioned at least one identified error pattern. KUs = knowledge units flagged in that week. Patterns = error patterns after Step 3.2 consolidation. The rightmost column shows the single most prevalent error pattern across both stages, with its per-pattern cohort coverage in parentheses (the percentage of groups whose feedback referenced that specific pattern). Weeks without data for a given stage are marked "-".

4.1.4. Cross-Year Analysis

Across the four cohorts from 2022 to 2025, the cross-year matching produced 46 pattern groups for Stage A, of which 27 appeared in two or more years and 15 persisted across three or more years. Stage B yielded 36 pattern groups, with 21 appearing in two or more years and 8 persisting across three or more years. Figure 6 tracks the five highest-coverage 2022 patterns across subsequent years.

As Fig. 6 shows, the five patterns follow different trajectories. **W5: Missing cross-section data and labeling in drawings**, the highest-coverage pattern in 2022, declines overall from 93.9% to 70.3% by 2025 after peaking at 97.3% in 2023. W4: **Missing justification for curve coordination** moves in the opposite direction, rising from 67.3% to 81.1% with a temporary dip in 2023. The remaining three patterns fluctuate without a consistent direction. These trajectories were presented to the teaching team for contextual interpretation.

4.2. Classification Validation (RQ2)

Table 6 presents the LLM screening agreement rates broken down by metric. Each metric (error type and knowledge unit relevance) reached 93–97% agreement across the five runs. Because a record was marked as confirmed only when all five runs agreed on both metrics, the overall confirmation rate was approximately 90% under each taxonomy. The remaining records, 166 (9.6%) under the data-driven taxonomy and 187 (10.8%) under the expert taxonomy, were flagged for human examination.

Table 6
LLM screening confirmation rates across five independent runs

Metric	Data-Driven Stage A	Data-Driven Stage B	Expert Stage A	Expert Stage B
Error type agreement	96.5%	95.0%	95.4%	93.1%
Knowledge unit relevance agreement	97.4%	95.6%	97.0%	95.5%

Table 7 summarizes the human review results. As the reviewed sample over-represents flagged records relative to the full dataset, agreement was estimated separately for flagged and non-flagged records within each error category, then weighted by category proportion of the full dataset to produce full-dataset estimates.

Although the LLM screening confirmation rates in Table 6 are comparable across taxonomies, the data-driven taxonomy exhibits substantially higher agreement with the human reviewer (89.1%, κ = 0.91) than the expert taxonomy (74.7%, κ = 0.64). The lower κ values on flagged cases (0.33 vs. 0.91 for the data-driven taxonomy; 0.21 vs. 0.64 for the expert taxonomy) indicate that the screening phase effectively concentrates disagreement with human judgment into the flagged subset. Agreement on knowledge unit relevance remains high for both taxonomies (92.0% vs. 89.7%).

Table 7
Human reviewer agreement on error category and knowledge unit assignments across both taxonomies

Metric	Data-Driven	Expert
Total records reviewed	266	302
Error type classification		
Flagged agreement	47.6% (79/166)	28.3% (53/187)
Non-flagged agreement	93.0% (93/100)	68.7% (79/115)
Flagged Cohen's kappa	0.33	0.21
Non-flagged Cohen's kappa	0.91	0.64
Full-dataset agreement (per-category weighted)	89.1	74.7
Knowledge unit relevance		
Flagged agreement	66.3% (110/166)	63.6% (119/187)
Non-flagged agreement	92.0% (92/100)	91.3% (105/115)
Full-dataset agreement (per-category weighted)	92.0%	89.7%

Table 8 disaggregates these full-dataset estimates by error category to identify where discrepancies are concentrated. For the data-driven taxonomy, four of the five categories achieved category-level agreement above 90%, with the remaining gap concentrated in **Procedural / Process Flaws** (65.8%), reflecting the difficulty of separating procedural errors from related categories of missing justification and reasoning gaps when a single piece of feedback touches on more than one. For the expert taxonomy, discrepancies are more pervasive and structural: **Careless** (21.4%) and **Procedural** (12.1%) produced agreement well below acceptable thresholds for content analysis, indicating that their conceptual boundaries could not be applied consistently. A further structural mismatch emerged outside the per-category figures: 102 records (5.9% of the dataset) were judged as belonging to a **Presentation / Communication** category absent from the expert taxonomy, forcing them into ill-fitting categories such as **Other** or **Omission**. The data-driven taxonomy, derived from the feedback corpus, included **Presentation / Communication** as a distinct category, removing this source of mismatch.

Table 8
Per-category agreement showing where pipeline annotations diverge from human judgment

Taxonomy	Error Category	Population	Category-Level Agreement
Data-driven	Presentation / Communication	6.3%	99.1%
	Positive	27.8%	94.8%
	Reasoning & Interpretation Errors	3.9%	91.7%
	Incomplete / Missing Content & Justification	48.0%	91.0%
	Procedural / Process Flaws	13.9%	**65.8%**
Expert	Positive	28.4%	98.0%
	Calculation	4.1%	79.9%
	Misconception	4.3%	73.2%
	Omission	55.2%	71.4%
	Careless	1.6%	**21.4%**
	Procedural	5.8%	**12.1%**
	Other	0.6%	-

Bold marks categories where discrepancies are most concentrated. Within each taxonomy, categories are ordered by descending agreement. The expert taxonomy's "Other" row is excluded from full-dataset estimates due to insufficient human review data (10 records, all flagged, 0% agreement). A further 102 records (5.9% of the dataset) were re-assigned by the reviewer to a Presentation / Communication category absent from the expert taxonomy.

Taken together, the discrepancy patterns suggest that alignment with human judgment depends less on whether a taxonomy is conceptually well-formed than on whether its category boundaries can be applied consistently to the feedback that markers actually produce. In the data-driven taxonomy, four of the five categories sustain agreement above 90%, including the dominant **Incomplete / Missing Content & Justification** (91.0% on 48.0% of records); only **Procedural / Process Flaws** (65.8%) falls noticeably below. In the expert taxonomy, by contrast, every non-Positive category falls below 90%, including the dominant **Omission** (71.4% on 55.2% of records), while **Careless** (21.4%) and **Procedural** (12.1%) collapse further. The expert taxonomy thus shows pervasive operationalization difficulties rather than a few problematic edge categories, whereas the data-driven taxonomy preserved the distinctions that emerged from the feedback corpus and produced more consistent agreement across categories of different sizes.

4.3. Teaching Team Review (RQ3)

Inductive thematic analysis of the focus-group transcript identified four themes, capturing how the teaching team moved from recognizing the scale of familiar errors, through reflecting on past teaching practice and proposing curriculum improvements, to identifying boundaries on the use of such evidence. Each theme is presented below with supporting verbatim quotes.

4.3.1. Scale of Recurring Errors

The teaching team's first engagement with the curriculum-aligned error patterns was a check against their own marking experience. All three participants confirmed that the types of errors identified by the proposed workflow were familiar and that the ranked priorities aligned with what they encountered during marking; the pipeline's quantified output and their experiential knowledge were seen as mutually validating.

The error categories themselves did not surprise the team. Their quantified scale, however, did. As TA1 reflected:

Before seeing this analysis, I really couldn't have imagined... I knew some students had problems with this, but I didn't expect it to be the mainstream issue.

UC expressed comparable surprise on specific topics:

The report says only 25% of students produced complete cross-section drawings and nearly two-thirds missed curve coordination, even after I had really emphasized it during class.

The proportion of students affected on topics the team believed they had thoroughly covered prompted them to ask why these errors persist despite existing instruction. UC reasoned that instruction is typically given weeks before students work on the relevant task, so "*by the time they're actually doing it, a lot of it has slipped from memory*". TA2 offered a complementary explanation, noting that "*some of them mostly go off the brief and don't always pick up on what we explain in class*". The pipeline did not identify these root causes directly; what it provided was a structured starting point from which the team could examine them.

4.3.2. Reflection on Past Teaching Adjustments

Beyond surfacing scale within a single year, the cross-year trend data allowed the team to look back and trace whether earlier teaching adjustments had corresponded to changes in student performance. UC recalled refining the 2025 project brief and providing more detailed in-class explanations with worked examples:

I made it (project brief) more detailed, specifying what should be done at each point... it's possible that clearer requirements actually made them follow through.

The pipeline data showed corresponding error reductions on the topics UC had targeted, which the team interpreted as consistent with these adjustments. The cross-year view also revealed a different kind of

finding, namely a mismatch between teaching effort and observed errors. Some topics receiving substantial teaching time showed consistently low error rates across cohorts, while others showed persistent or worsening problems. UC observed:

I spent a lot of effort on some of them (topics), and they do have basically no major issues.

This contrast prompted the team to question whether teaching effort had been allocated proportionally to where students actually struggle.

4.3.3. Curriculum Improvement Actions

The reflections in the previous theme led directly to concrete proposals for curriculum changes. UC outlined three improvements anchored in the identified error patterns: "provide a benchmark to show students what a good example looks like", "in-class demonstration" to walk through problems that "closely mimic the actual assessment task", and "weekly alerts to let students know that is the errors they may have".

The data also prompted participants to reconsider how teaching hours are allocated across topics. Topics currently occupying multiple weeks of instruction that "have basically no major issues" in the pipeline data were contrasted with those showing persistently high error rates. UC concluded that time should be shifted from the former to the latter:

Economic evaluation in practice is really complex and may need more time. I used to spend one week, now I probably need to spend two weeks on it.

These proposals illustrate how the curriculum-aligned format of the output, organized by topic and teaching week, supported the team in moving from recognizing patterns to specifying changes at the level of individual topics and time allocations.

4.3.4. Limits of the Evidence

While participants found the patterns useful, they were also clear about what the evidence cannot do and how it should not be used. TA1 observed that the pipeline's error patterns cannot account for how the course is actually delivered:

Content like oral communication, in-class demonstrations, and help-desk, they probably have no trace in the written materials that the pipeline analyzes. So that the results may be most useful to those who already know what has been taught.

TA2 added that the output should not be treated as standalone evidence:

The teaching team would look at it first. They know what is going on and won't treat it as the only evidence, but more like a reference.

Participants also cautioned that cross-year improvements do not always reflect genuine learning gains. The adoption of computer-aided design software in 2025, for instance, improved drawing quality but may represent a superficial improvement rather than understanding of professional drawing standards. The strongest caution concerned the use of the output for evaluating teaching: as UC put it, echoed by TA1, "this definitely cannot be used for teaching evaluation". Using error rates as an external measure of teaching quality could obscure ongoing instructional efforts and discourage the experimentation that drives improvement.

5. Discussion

The findings reposition assessment feedback. Where feedback has predominantly been framed as a student-facing communication channel (Boud & Molloy, 2013; Carless & Boud, 2018), this study shows that the same text, when aggregated and aligned with curriculum content, surfaces the specific concepts and procedures students are getting wrong at a scale markers experience individually but cannot quantify. Viewed this way, feedback becomes a diagnostic data source for teaching, complementing rather than replacing its student-facing function, and broadens the empirical base for teaching-improvement decisions beyond what current teacher-facing analytics tools provide. The approach also extends the principle of constructive alignment (Biggs et al., 2022) beyond design-stage coherence between learning outcomes, teaching, and assessment: by tracing observed student errors back to specific taught content, it makes alignment failures empirically visible at the level where teaching adjustments can be made.

The data-driven taxonomy achieved substantially higher agreement with human judgment than the expert taxonomy (89.1% versus 74.7%). This makes a concrete empirical claim about category design for LLM-based classification of educational text: conceptual coherence in expert-designed categories is not a sufficient predictor of LLM-human agreement. The per-category analysis locates the gap in specific structural features of the expert design rather than in surface wording. Two of its categories (**Careless** and **Procedural**) produced agreement well below thresholds typically considered acceptable, and 5.9% of records were re-assigned by the human reviewer to a **Presentation / Communication** category entirely absent from the expert taxonomy. The expert taxonomy was conceptually well-formed but did not reflect the distinctions markers actually drew when writing feedback. This relocates the question of category design from how categories are worded, the focus of prior work on LLM sensitivity to definitional changes (Dunivin, 2025; Halterman & Keith, 2026), to where they are generated from: deriving categories from the corpus to be classified provides a more reliable basis than expert pre-specification. Future work could test whether this finding generalizes to other educational-text classification tasks where expert taxonomies are dominant, such as misconception detection or written-response coding.

A further contribution concerns how teaching teams actually engage with the resulting evidence. Participants did not treat the quantified patterns as standalone findings or prescriptive recommendations. Instead, they triangulated the data with contextual information the pipeline cannot capture, including classroom demonstrations, project-brief revisions, help-desk interactions, and their

tacit reading of cohort behavior, in order to interpret why specific errors persisted or declined. This interpretive process generated concrete curriculum decisions, such as reallocating teaching hours toward topics with persistent errors and introducing worked examples at specific points in the schedule. It also generated explicit cautions: cross-year improvements may not reflect genuine learning gains, written materials cannot represent oral instruction, and error rates should not be repurposed as a measure of teaching quality. The pattern is consistent with arguments that effective instructional decisions draw on a combination of structured evidence and contextual knowledge rather than on data alone (Mandinach & Schildkamp, 2021), and with situated models of instructor analytics use that frame the value of analytics as residing in educators' interpretive work (Wise & Jung, 2019). The contribution here is to extend these models empirically from the engagement and performance metrics on which they were originally based to curriculum-aligned content evidence derived from assessment feedback, and to show that teaching teams treat such evidence as a starting point for reflection rather than as a replacement for their professional judgment.

6. Limitations

The case study was conducted within a single engineering course at one institution, using group-based project reports as the assessment format, and the taxonomy was generated from one year of data and applied to subsequent years. Although the method is designed to be domain-independent, requiring only feedback and teaching materials as inputs, its effectiveness across other disciplines, assessment formats such as individual assignments or examinations, and institutional contexts remains to be tested. Future work should evaluate the method in these broader settings and examine when taxonomy regeneration becomes necessary.

The knowledge base was constructed from written teaching materials and therefore does not represent oral emphasis, in-class demonstrations, or help-desk interactions that also shape student learning. Relying solely on written materials means the curriculum-alignment step may not fully reflect the instruction delivered in the course, and the resulting error patterns should be interpreted by the teaching team. Similarly, changes in error coverage across years may reflect not only instructional adjustments but also differences in cohort composition or marking practice, and the method cannot distinguish among these factors. Future work could explore integrating non-text teaching data and adopting longitudinal designs that control for confounding variables.

Within a single course, the pool of staff with sufficient marking experience and familiarity with the content is inherently small, which constrained validation to a limited number of reviewers and made independent cross-validation between experienced reviewers infeasible. The classification validation therefore relied on a single human reviewer with three years of marking experience, and the teaching team review involved three participants covering all four cohorts. To make the classification validation tractable, the reviewer was shown the pipeline's linked knowledge unit during assessment, and this partial blinding may have primed category judgments toward error types typically associated with the visible knowledge unit. While these participants provided informed and consistent judgments, validation

with multiple independent reviewers, fully blinded annotation protocols, and teaching teams from different courses would further strengthen confidence in the method's reliability and the generalizability of its output.

7. Conclusion

This paper presented an LLM-based method that transforms unstructured assessment feedback into curriculum-aligned error analysis through three steps: data-driven taxonomy generation, retrieval-augmented curriculum mapping, and multi-level aggregation into quantified error patterns organized by teaching sequence. In a case study of 6072 pieces of feedback across four cohorts of an undergraduate engineering course, the data-driven taxonomy achieved 89.1% classification agreement with human judgment, outperforming an expert taxonomy (74.7%), and the teaching team confirmed the output as consistent with their marking experience and actionable for curriculum improvement. By treating assessment feedback as a data source for teaching improvement, this work proposes a novel method for systematic use of qualitative educational data to inform curriculum decisions.

Declarations

Author Contribution

L.H. initiated the study, developed the methodology, implemented the LLM-based analysis pipeline, performed the experiments and data analysis, and wrote the original draft. J.X. contributed to the study design and led the qualitative analysis of the focus group data. L.G. curated the feedback dataset and produced the human-coded reference used to validate the annotations. N.Z. provided resources and contributed to project supervision. Q.S. and D.T. contributed to the conceptualization of the study and to the implementation of the retrieval-augmented generation component. Y.L. provided resources and supervised the project. All authors reviewed and revised the manuscript.

Data Availability

The data that supports the findings of this study are not publicly available due to restrictions imposed by the approval granted by Monash University, which permits use of the data solely for the purposes of this study. The data may be made available from the corresponding author upon reasonable request and with the permission of Monash University.

References

1. Aldino, A. A., Tsai, Y. S., Gupte, S., Henderson, M., Nath, D., Gašević, D., & Chen, G. (2025). Analytics of learner-centered feedback: A large-scale case study in higher education. *Computers & Education*, *237*, 105360. https://doi.org/10.1016/j.compedu.2025.105360

2. Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2025). Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, *8*, 17. https://doi.org/10.1007/s42001-024-00345-9
3. Arafeh, S. (2016). Curriculum mapping in higher education: A case study and proposed content scope and sequence mapping tool. *Journal of Further and Higher Education*, *40*(5), 585–611. https://doi.org/10.1080/0309877X.2014.1000278
4. Baartman, L. K. J., & Quinlan, K. M. (2024). Assessment and feedback in higher education reimagined: Using programmatic assessment to transform higher education. *Perspectives: Policy and Practice in Higher Education*, *28*(2), 57–67. https://doi.org/10.1080/13603108.2023.2283118
5. Banihashem, S. K., Gašević, D., Noroozi, O., Jarodzka, H., Joosten-ten Brinke, D., & Drachsler, H. (2025). Optimizing formative assessment with learning analytics. *Review of Educational Research*, *95*(1), 1–49. https://doi.org/10.3102/00346543251370753
6. Banihashem, S. K., Noroozi, O., van Ginkel, S., Macfadyen, L. P., & Biemans, H. J. A. (2022). A systematic review of the role of learning analytics in enhancing feedback practices in higher education. *Educational Research Review*, *37*, 100489. https://doi.org/10.1016/j.edurev.2022.100489
7. Biggs, J., Tang, C., & Kennedy, G. (2022). *Teaching for quality learning at university* (5th ed.). McGraw-Hill Education.
8. Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102
9. Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education*, *41*(3), 466–481. https://doi.org/10.1080/02602938.2015.1024607
10. Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, *38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462
11. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa
12. Buckingham Shum, S., Ferguson, R., & Martinez-Maldonado, R. (2019). Human-centred learning analytics. *Journal of Learning Analytics*, *6*(2), 1–9. https://doi.org/10.18608/jla.2019.62.1
13. Cai, C., Hong, S., Ma, M., Feng, H., Du, S., Chow, M., Teo, W. L., Liu, S., & Fan, X. (2025). Analyzing the teaching and learning environments through student feedback at scale: A multi-agent LLMs framework. *Education and Information Technologies*, *30*, 21815–21847. https://doi.org/10.1007/s10639-025-13633-2
14. Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354
15. Carlson, N. A., & Burbano, V. (2026). The use of LLMs to annotate data in management research: Foundational guidelines and warnings. *Strategic Management Journal*, *47*(3), 699–725.

https://doi.org/10.1002/smj.70023

16. Cavalcanti, A. P., Diego, A., Mello, R. F., Mangaroska, K., Nascimento, A., Freitas, F., & Gašević, D. (2020). How good is my feedback? A content analysis of written feedback. In *Proceedings of the 10th International Learning Analytics and Knowledge Conference (LAK '20)* (pp. 428–437). ACM. https://doi.org/10.1145/3375462.3375477
17. Claassen, A., Mirriahi, N., Kovanović, V., & Dawson, S. (2025). From data to design: Integrating learning analytics into educational design for effective decision-making. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)* (pp. 558–567). ACM. https://doi.org/10.1145/3706468.3706541
18. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104
19. Dai, S. C., Xiong, A., & Ku, L. W. (2023). LLM-in-the-loop: Leveraging large language model for thematic analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 9993–10001). ACL. https://doi.org/10.18653/v1/2023.findings-emnlp.669
20. Derham, C., Balloo, K., & Winstone, N. (2022). The focus, function and framing of feedback information: Linguistic and content analysis of in-text feedback comments. *Assessment & Evaluation in Higher Education*, *47*(6), 896–909. https://doi.org/10.1080/02602938.2021.1969335
21. Dunivin, Z. O. (2025). Scaling hermeneutics: A guide to qualitative coding with LLMs for reflexive content analysis. *EPJ Data Science*, *14*(1). Article 28. https://doi.org/10.1140/epjds/s13688-025-00548-8
22. Echeverria, V., Martinez-Maldonado, R., Shum, B., Chiluiza, S., Granda, K., R., & Conati, C. (2018). Exploratory versus explanatory visual learning analytics: Driving teachers' attention through educational data storytelling. *Journal of Learning Analytics*, *5*(3), 73–97. https://doi.org/10.18608/jla.2018.53.6
23. Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T. S., & Li, Q. (2024). A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)* (pp. 6491–6501). ACM. https://doi.org/10.1145/3637528.3671470
24. Fung, S. C. E., Wong, M. F., & Tan, C. W. (2024). Automatic feedback generation on K-12 students' data science education by prompting cloud-based large language models. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24)* (pp. 255–258). ACM. https://doi.org/10.1145/3657604.3664673
25. Gottipati, S., Shankararaman, V., & Lin, J. R. (2018). Text analytics approach to extract course improvement suggestions from students' feedback. *Research and Practice in Technology Enhanced Learning*, *13*., Article 6. https://doi.org/10.1186/s41039-018-0073-0
26. Halterman, A., & Keith, K. A. (2026). Codebook LLMs: Evaluating LLMs as Measurement Tools for Political Science Concepts. *Political Analysis*, *34*(2), 188–204. https://doi.org/10.1017/pan.2025.10017

27. Ho, J. E. (2022). What counts? The critical role of qualitative data in teachers' decision making. *Evaluation and Program Planning*, *91*., Article 102046. https://doi.org/10.1016/j.evalprogplan.2021.102046
28. Istencioglu, T., Boud, D., Dawson, P., & Yang, L. (2026). Teacher feedback literacy as an emerging field: A scoping review. *Assessment & Evaluation in Higher Education*, 1–20. https://doi.org/10.1080/02602938.2026.2631533
29. Jacobs, S., & Jaschke, S. (2024). Leveraging lecture content for improved feedback: Explorations with GPT-4 and retrieval augmented generation. In *2024 36th International Conference on Software Engineering Education and Training (CSEE&T)* (pp. 1–5). IEEE. https://doi.org/10.1109/CSEET62301.2024.10663001
30. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274
31. Klein, C., Lester, J., Rangwala, H., & Johri, A. (2019). Technological barriers and incentives to learning analytics adoption in higher education: Insights from users. *Journal of Computing in Higher Education*, *31*(3), 604–625. https://doi.org/10.1007/s12528-019-09210-5
32. Kökver, Y., Pektaş, H. M., & Çelik, H. (2025). Artificial intelligence applications in education: Natural language processing in detecting misconceptions. *Education and Information Technologies*, *30*, 3035–3066. https://doi.org/10.1007/s10639-024-12919-1
33. Krippendorff, K. (2019). *Content analysis: An introduction to its methodology*. SAGE Publications, Inc.. https://doi.org/10.4135/9781071878781
34. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* (pp. 9459–9474). Curran Associates. https://dl.acm.org/doi/abs/10.5555/3495724.3496517
35. Li, X., Henriksson, A., Duneld, M., Nouri, J., & Wu, Y. (2024). Supporting Teaching-to-the-Curriculum by Linking Diagnostic Tests to Curriculum Goals: Using Textbook Content as Context for Retrieval-Augmented Generation with Large Language Models. In A. M. Olney, I. A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial Intelligence in Education. AIED 2024. Lecture Notes in Computer Science* (Vol. 14829). Springer. https://doi.org/10.1007/978-3-031-64302-6_9
36. Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, *8*, 100417. https://doi.org/10.1016/j.caeai.2025.100417
37. Lin, J., Dai, W., Lim, L. A., Tsai, Y. S., Mello, R. F., Khosravi, H., Gašević, D., & Chen, G. (2023). Learner-centred analytics of feedback content in higher education. In *Proceedings of the 13th International*

Learning Analytics and Knowledge Conference (LAK '23) (pp. 100–110). ACM. https://doi.org/10.1145/3576050.3576064

38. Mandinach, E. B., & Schildkamp, K. (2021). Misconceptions about data-based decision making in education: An exploration of the literature. *Studies in Educational Evaluation*, *69*. https://doi.org/10.1016/j.stueduc.2020.100842. Article 100842.
39. McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664
40. Mello, R. F., Neto, R., Fiorentino, G., Alves, G., Arêdes, V., Silva, J. V. G. F., Falcão, T. P., & Gašević, D. (2022). Enhancing instructors' capability to assess open-response using natural language processing and learning analytics. In I. Hilliger, P. J. Muñoz-Merino, T. De Laet, A. Ortega-Arranz, & T. Farrell (Eds.), *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption* (LNCS, Vol. 13450, pp. 102–115). Springer International Publishing. https://doi.org/10.1007/978-3-031-16290-9_8
41. Milesi, M. E., & Martinez-Maldonado, R. (2024). Data storytelling in learning analytics? A qualitative investigation into educators' perceptions of benefits and risks. In *Proceedings of the 14th International Learning Analytics and Knowledge Conference (LAK '24)* (pp. 476–486). ACM. https://doi.org/10.1145/3636555.3636865
42. Nafa, F., Khan, J. I., Othman, S., & Babour, A. (2016). Discovering Bloom taxonomic relationships between knowledge units using semantic graph triangularity mining. In *Proceedings of the 2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)* (pp. 224–231). IEEE. https://doi.org/10.1109/CyberC.2016.52
43. Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090
44. Osakwe, I., Chen, G., Whitelock-Wainwright, A., Gašević, D., Cavalcanti, A. P., & Mello, R. F. (2022). Towards automated content analysis of educational feedback: A multi-language study. *Computers and Education: Artificial Intelligence*, *3*, 100059. https://doi.org/10.1016/j.caeai.2022.100059
45. Panadero, E., Jonsson, A., Pinedo, L., & Fernandez-Castilla, B. (2023). Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: A meta-analytic review. *Educational Psychology Review*, *35*, 113. https://doi.org/10.1007/s10648-023-09823-4
46. Parfenova, A., Marfurt, A., Pfeffer, J., & Denzler, A. (2025). Text annotation via inductive coding: Comparing human experts to LLMs in qualitative data analysis. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 6471–6484). ACL. https://doi.org/10.18653/v1/2025.findings-naacl.361
47. Pereira, E., Nsair, S., Pereira, R., L., & Grant, K. A. (2024). Constructive alignment in a graduate-level project management course: An innovative framework using large language models. *International*

Journal of Educational Technology in Higher Education, *21*, 25. https://doi.org/10.1186/s41239-024-00457-2

48. Phung, D. V., & Michell, M. (2022). Inside teacher assessment decision-making: From judgement gestalts to assessment pathways. *Frontiers in Education*, *7*, 830311. https://doi.org/10.3389/feduc.2022.830311
49. Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714
50. Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, *61*(3), 257–273. https://doi.org/10.1080/00131881.2019.1625716
51. Shermis, M. D., & Wilson, J. (Eds.). (2024). *The Routledge International Handbook of Automated Essay Evaluation (1st ed.)*. Routledge. https://doi.org/10.4324/9781003397618
52. Shimizu, Y., & Kang, H. (2025). Research on classroom practice and students' errors in mathematics education: A scoping review of recent developments for 2018–2023. *ZDM - Mathematics Education*, *57*, 695–710. https://doi.org/10.1007/s11858-025-01704-0
53. Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. https://doi.org/10.3102/0013189X015002004
54. Than, N., Fan, L., Law, T., Nelson, L. K., & McCall, L. (2025). Updating The Future of Coding: Qualitative coding with generative large language models. *Sociological Methods & Research*, *54*(3), 849–888. https://doi.org/10.1177/00491241251339188
55. Vanlommel, K., Van Gasse, R., Vanhoof, J., & Van Petegem, P. (2017). Teachers' decision-making: Data based or intuition driven? *International Journal of Educational Research*, *83*, 75–83. https://doi.org/10.1016/j.ijer.2017.02.013
56. Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98–110. https://doi.org/10.1016/j.chb.2018.07.027
57. Vikan, M., Aryan, R., Kannelønning, M. S., Riegler, M. A., & Danielsen, S. O. (2026). Reflecting on LLM support in reflexive thematic analysis: An exploratory study. *Qualitative Health Research*, *36*(2–3), 191–205. https://doi.org/10.1177/10497323251365211
58. Wan, M., Safavi, T., Jauhar, S. K., Kim, Y., Counts, S., Neville, J., Suri, S., Shah, C., White, R. W., Yang, L., Andersen, R., Buscher, G., Joshi, D., & Rangan, N. (2024). TnT-LLM: Text mining at scale with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 5836–5847). Association for Computing Machinery. https://doi.org/10.1145/3637528.3671647
59. Wise, A. F., & Jung, Y. (2019). Teaching with analytics: Towards a situated model of instructional decision-making. *Journal of Learning Analytics*, *6*(2), 53–69. https://doi.org/10.18608/jla.2019.62.4
60. Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

61. Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Figures

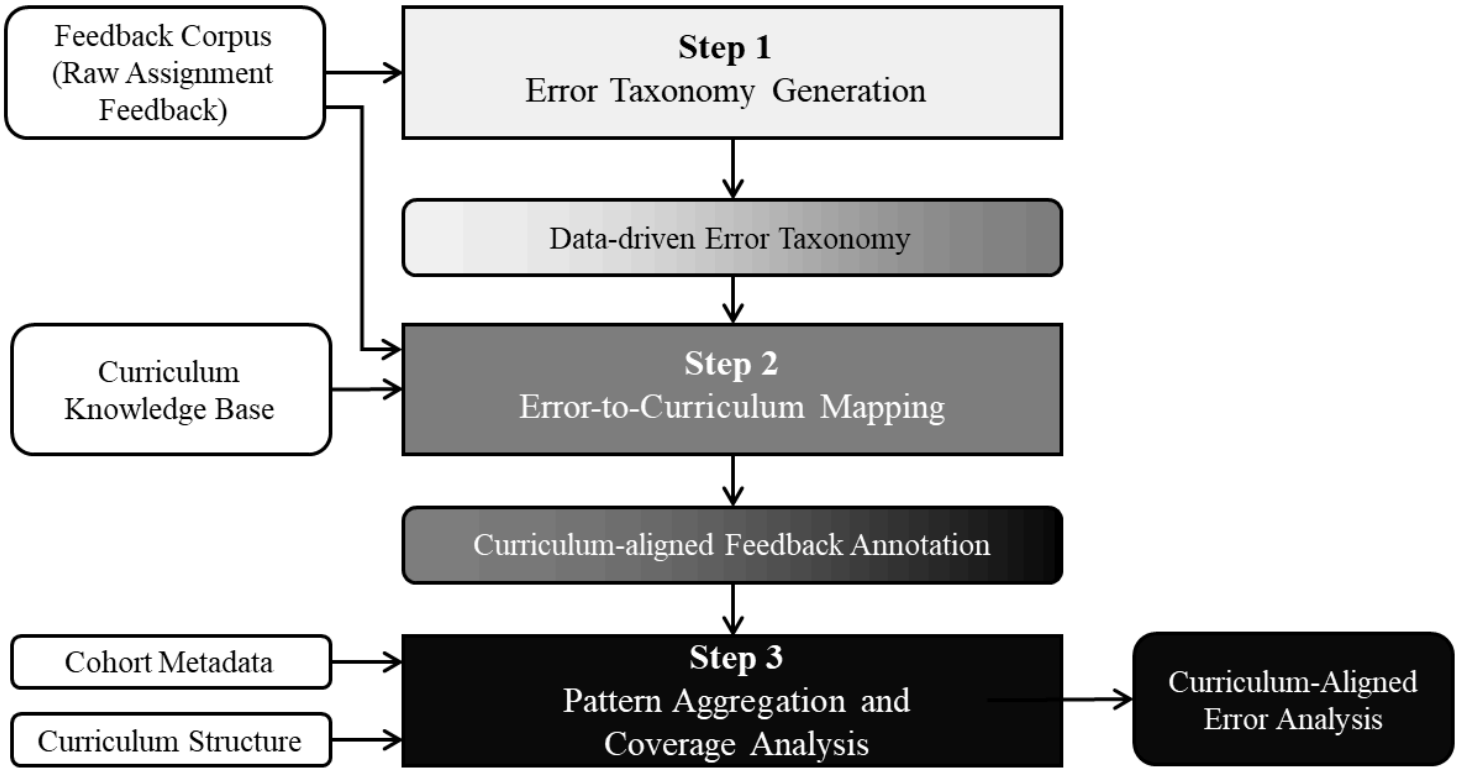


Figure 1

Overview of the three-step workflow for curriculum-aligned error analysis.

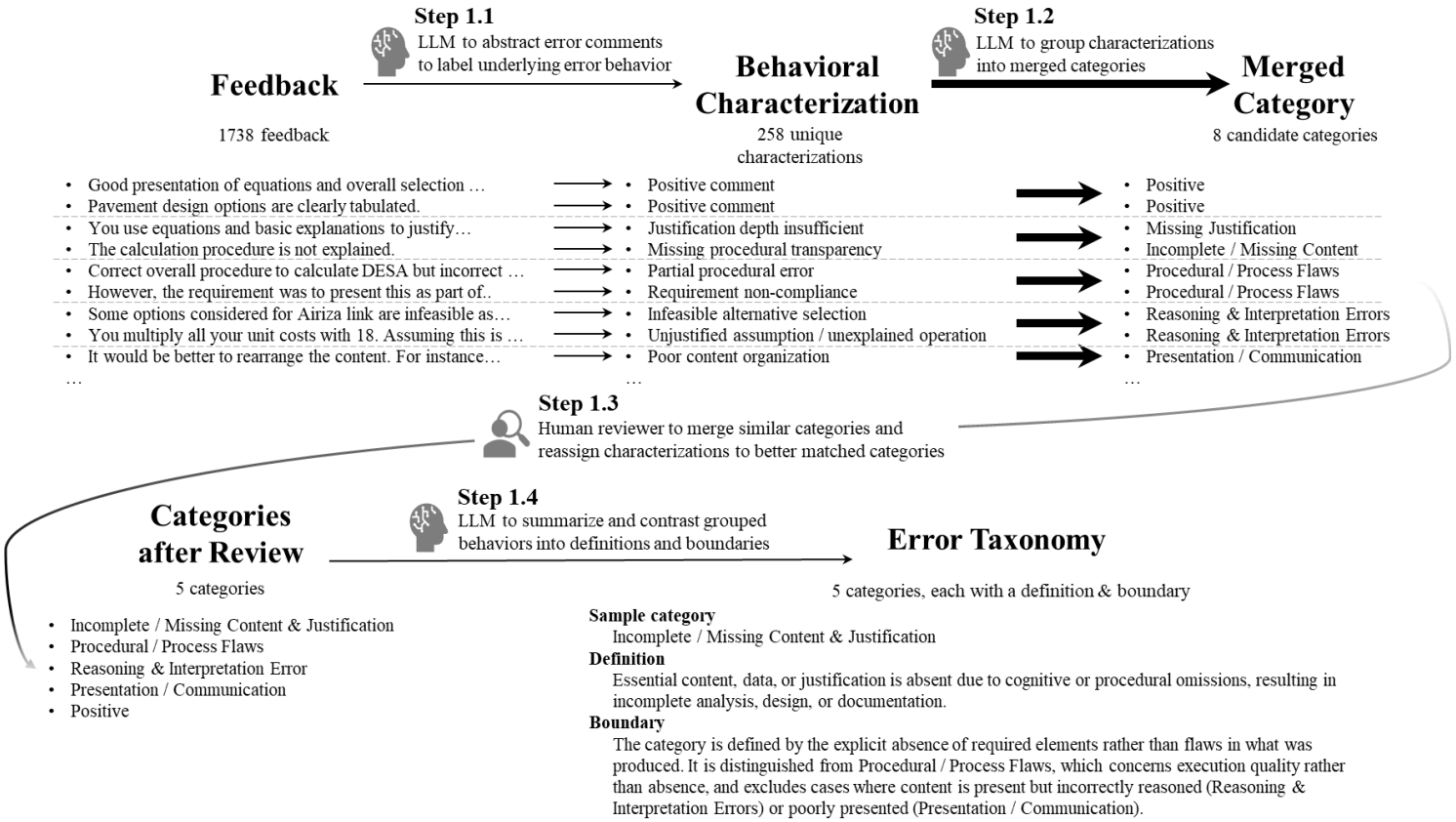


Figure 2

Error taxonomy generation workflow.

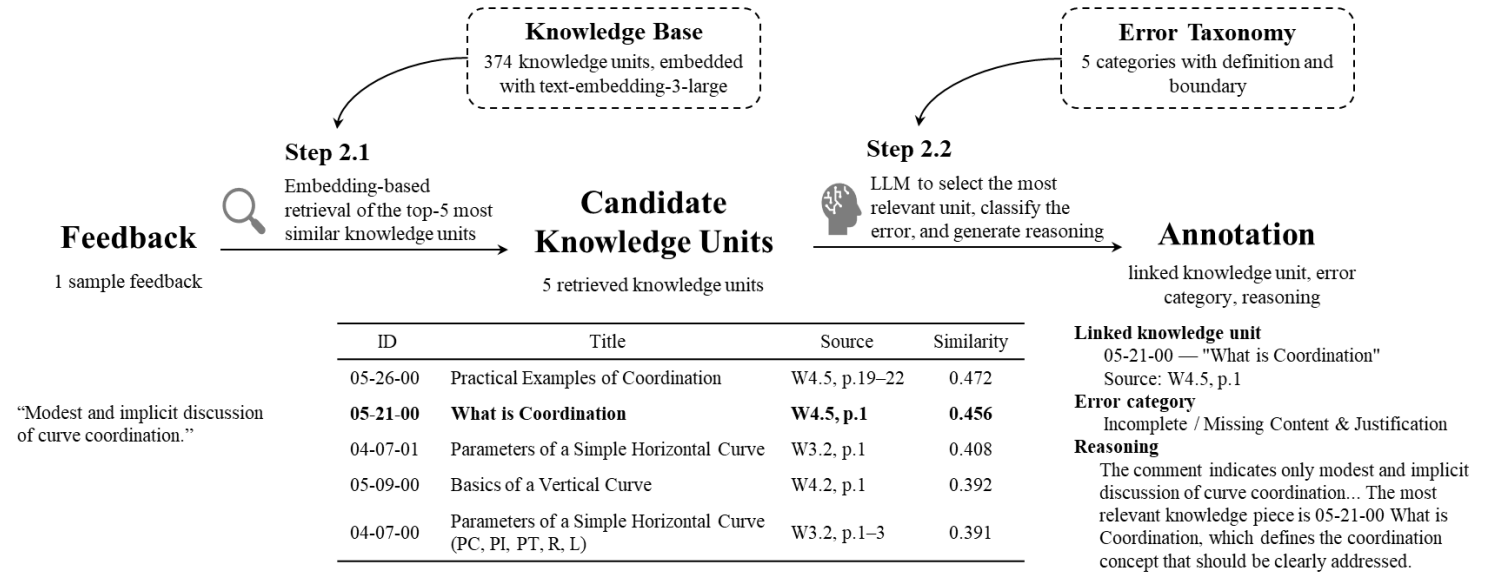


Figure 3

Error-to-curriculum mapping process, illustrated with one sample feedback. The knowledge unit selected by the LLM in Step 2.2 is shown in bold in the candidate list.

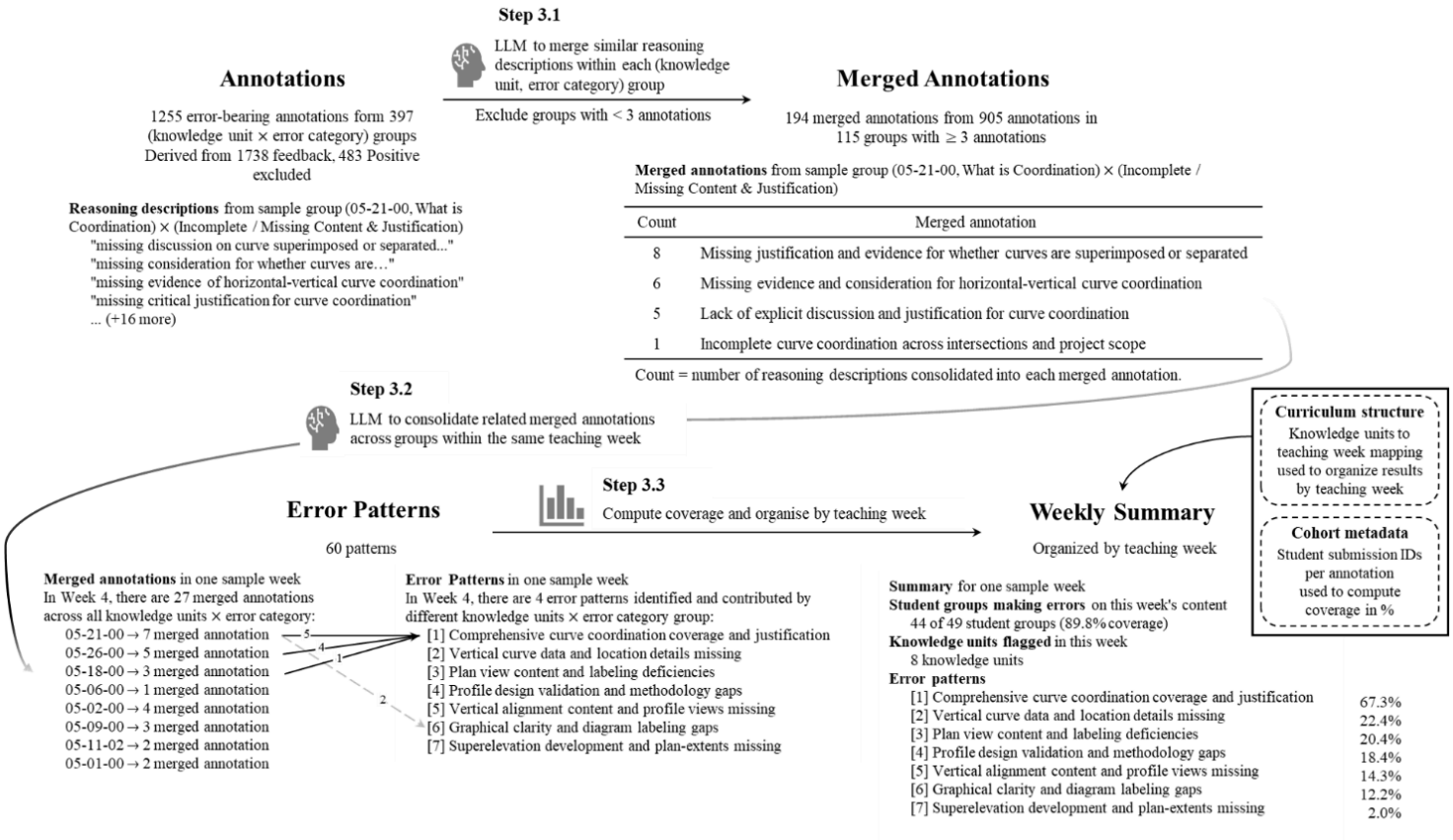


Figure 4

Pattern aggregation and coverage analysis workflow, illustrated with Week 4 of the 2022 dataset

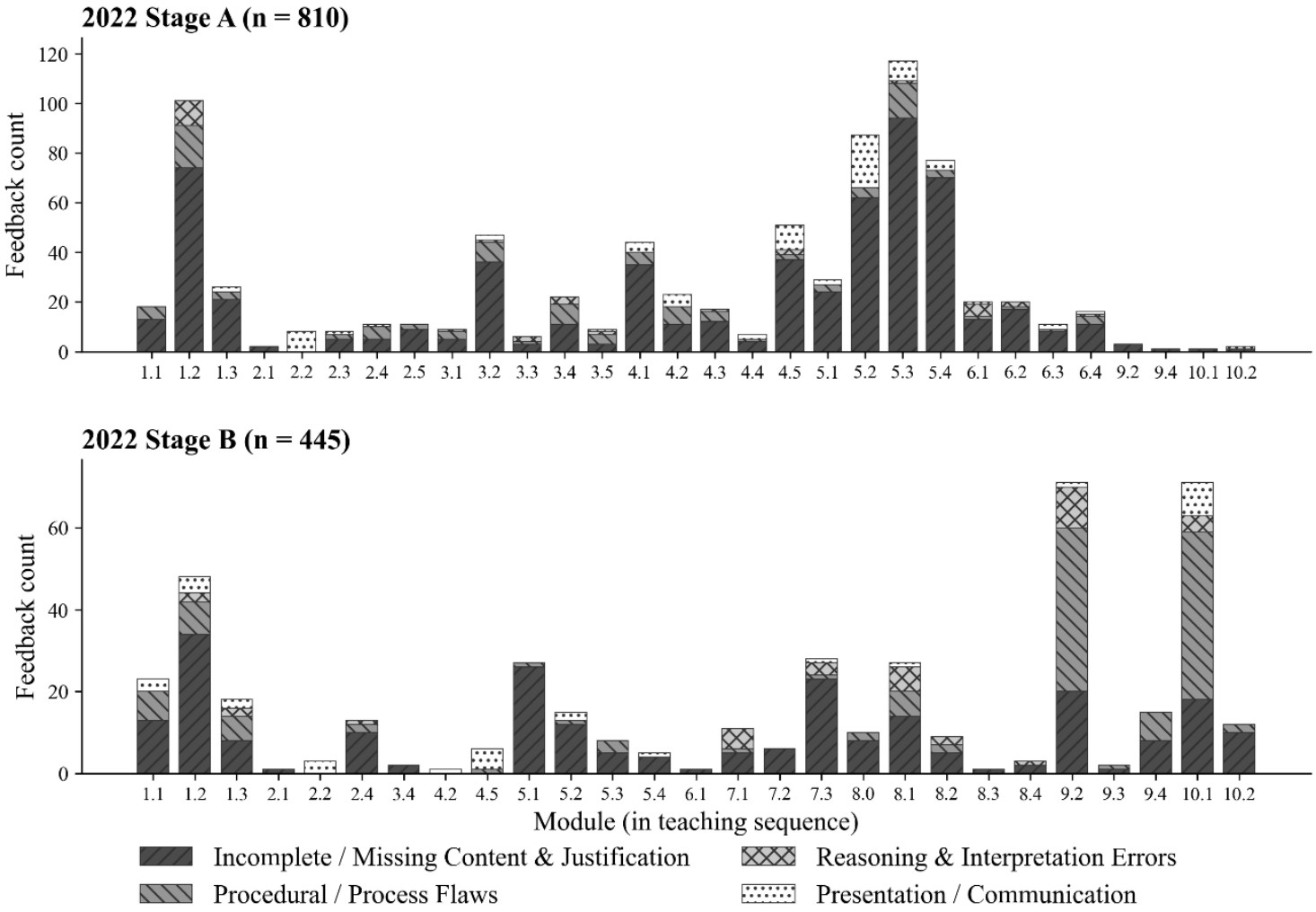


Figure 5

Distribution of error-bearing feedback across curriculum modules for Stage A (top) and Stage B (bottom), organized by teaching sequence. Each bar represents a slide module, with stacked segments indicating error category composition.

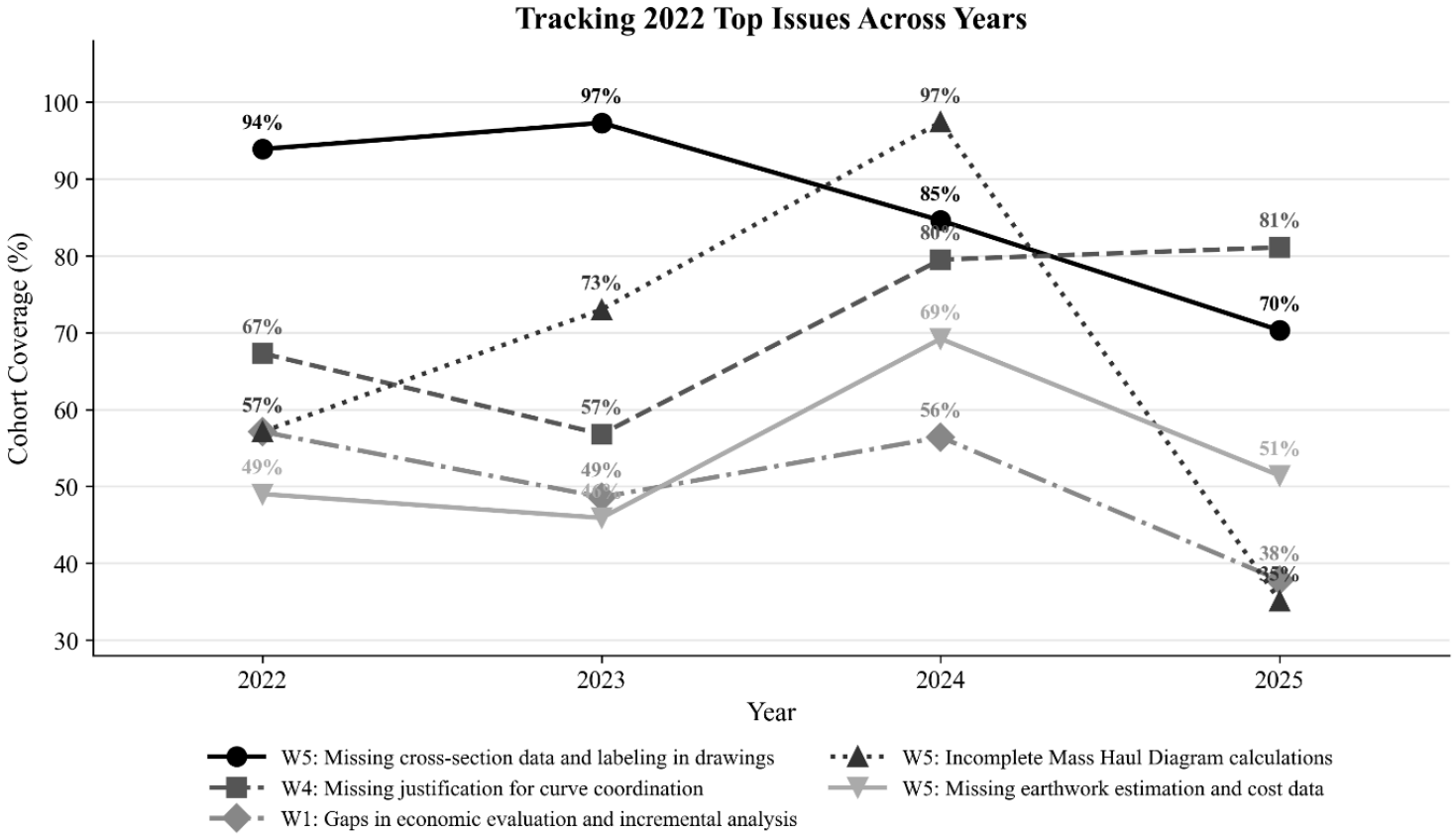


Figure 6

Cohort coverage of the five most prevalent error patterns identified in 2022 across the two stages, tracked across 2022–2025. Coverage indicates the percentage of student groups in a cohort whose feedback referenced the corresponding curriculum-aligned error pattern.