

# Research on Teaching of Structural Safety Judgment for Railway Vehicles Driven by Controlled Errors of Generative Artificial Intelligence

Yiliang Shu<sup>1\*</sup>, Daoyun Chen<sup>1</sup>, Changjun Deng<sup>1</sup>, Dong Liu<sup>1</sup>, Xing Chen<sup>2</sup>, Jie Su<sup>3</sup>

1. School of Mechatronics and Vehicle Engineering, East China Jiaotong University, Nanchang, Jiangxi, 330199, China

2. School of Education, Jiangxi Institute of Applied Science and Technology, Nanchang, Jiangxi, 330100, China.

3. College of Mechanical and Vehicle Engineering, Changsha University of Science and Technology, Changsha, 410114, China

*\*Corresponding author: Yiliang Shu, ylshu@ecjtu.edu.cn*

**Copyright:** 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY-NC 4.0), permitting distribution and reproduction in any medium, provided the original author and source are credited, and explicitly prohibiting its use for commercial purposes.

**Abstract:** Aiming at the common problems in railway vehicle structure courses: students tend to overemphasize calculation results while ignoring applicable conditions, prioritize software operation over evidence verification, and easily develop judgment dependence after the introduction of generative artificial intelligence, this paper proposes a teaching framework for structural safety judgment driven by controlled errors. Taking welded joints of bogie frames, box structural responses under wheel polygon excitation, and local stress interpretation as research carriers, a case generation mechanism integrating retrieval-augmented generation, targeted error mutation, independent program verification and instructor review is designed. Teaching activities include independent preliminary judgment, evidence checking, revision and argumentation, and AI-free transfer evaluation. Teaching examples demonstrate that incorrectly substituting stress amplitude into stress-range-based fatigue curves can lead to an 87.5% underestimation of damage in this case. Under the specified single-mode near-resonance condition, the estimation of responses using a constant dynamic load factor is significantly low. The above results are used to characterize error consequences and do not represent classroom teaching effectiveness. Furthermore, an evaluation rubric covering error detection, correct adoption of information, evidence sufficiency and confidence calibration, together with three groups of control schemes, is put forward. This study develops reproducible professional cases and implementable evaluation routes, providing instructional design references for cultivating engineering judgment capabilities of railway vehicle engineers in the generative AI era. The learning benefits and applicable boundaries still need to be validated by real classroom data.

**Keywords:** Generative Artificial Intelligence; Controlled Errors; Railway Vehicles; Structural Fatigue; Engineering Judgment; Evidence Verification

**Published:** Oct 4, 2026

**DOI:** <https://doi.org/10.62177/jetp.v3i4.1744>

## 1. Introduction

Structural safety analysis of railway vehicles requires the correlation of service loads, structural responses, material or joint resistance, and assessment conditions. For bogie frame welded joints, axle interference-fit regions and gearbox housings,

identical stress values may correspond to different extraction methods, load histories and failure mechanisms. When course assessment focuses predominantly on final calculation results, students tend to neglect statements on model assumptions, data sources and the bounds of conclusions, leading to a disconnect between “being able to compute” and “being able to judge accurately”.

Generative artificial intelligence can rapidly assemble analysis procedures, explain terminology and generate computational scripts. Nevertheless, it may produce seemingly well-formulated and coherently reasoned recommendations embedded with critical errors. Bastani et al.<sup>[1]</sup> found, in a high-school mathematics context, that performance gains with tool support do not necessarily translate into learning benefits after tool removal. Their findings cannot be directly extrapolated to railway vehicle courses, yet they suggest that teaching evaluation should distinguish immediate task performance from standalone competence. For structural safety judgment, the core competency to be trained lies in students’ ability to verify evidence, identify inapplicable conditions, and take responsibility for interpreting conclusions.

Research on error-based learning emphasizes that errors can be converted into learning resources only through effective feedback and root-cause analysis<sup>[2]</sup>. Error cases in engineering education additionally need to be localizable, recalculable and rectifiable. Randomly collected erroneous responses from large language models make it difficult to control conceptual difficulty, the number of errors and their consequences; students may even guess answers merely from linguistic style. It is therefore necessary to combine instructor-vetted error mechanisms with the variant generation capability of generative AI to develop controlled teaching materials.

This paper addresses three research questions: how to convert key misjudgements in structural fatigue analysis into controllable cases; how to organize classroom learning via evidence verification and fading feedback; and how to evaluate students’ capabilities of independent judgement and reasonable evidence adoption on unfamiliar tasks. The contributions of this work lie in establishing a taxonomy of domain-specific errors and a paired-case generation workflow, providing recalculable analysis of error consequences, and constructing an evaluation framework accounting for missed detections, false judgements and evidence quality. This study does not rely on algorithmic novelty to substitute pedagogical innovation, nor does it present the proposed workflow as a fully implemented system.

## 2. Connotation and Teaching Mechanism of Structural Safety Judgment

### 2.1 From Calculation Correctness to Conclusion Justifiability

In this paper, structural safety judgment is defined as the capability to select rational analysis methods, examine calculation procedures, and clarify the supports and limitations of structural conclusions under given evidence and applicable conditions. The judgment target focuses on whether the available analysis can support the drawn conclusions, rather than requiring students to directly determine the safety of physical structures with incomplete data. Three judgment categories are adopted: conclusion justifiable, conclusion requiring revision, and insufficient evidence, to distinguish definite errors from unsubstantiated claims.

*Table 1. Competencies of Structural Safety Judgment and Observable Learning Evidence*

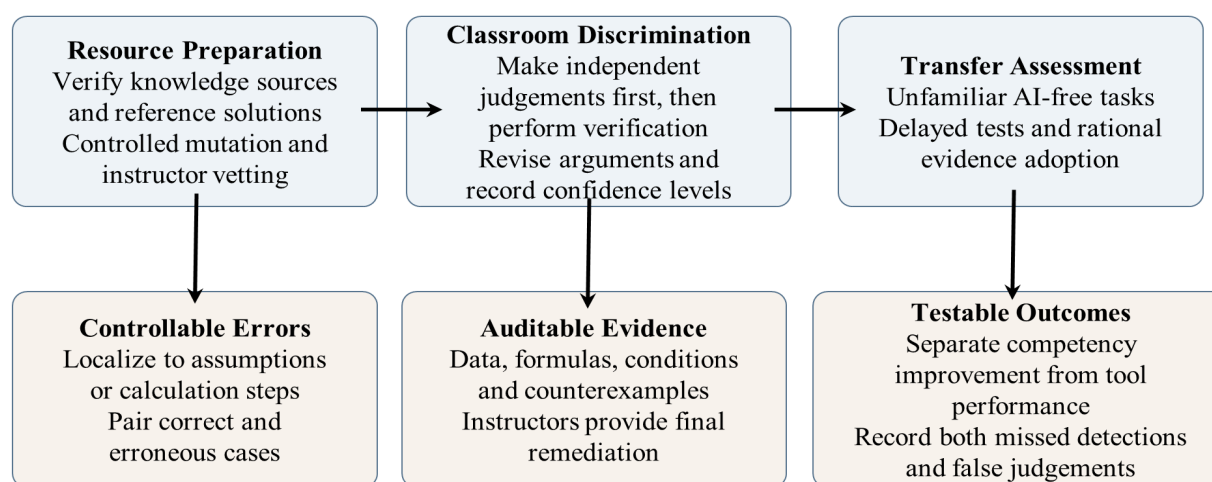
| Competency Dimension     | Professional Question                                                       | Observable Evidence                                                         |
|--------------------------|-----------------------------------------------------------------------------|-----------------------------------------------------------------------------|
| Condition identification | Does the stress definition match the fatigue curve?                         | List loads, stress metrics and applicable conditions of the curve           |
| Model review             | Does the static approximation capture dominant dynamic responses?           | Specify frequency range, mode participation and rationale for approximation |
| Evidence verification    | Do figures, data and citations support the conclusion?                      | Locate raw data or derivations and perform independent recalculation        |
| Conclusion revision      | Do changes in calculations alter safety judgments?                          | Quantify error consequences and draw bounded conclusions                    |
| Transfer reflection      | Does the method remain applicable for alternative components or load cases? | Independently select verification methods and account for uncertainties     |

## 2.2 Controlled Errors and Cognitive Conflict

Controlled errors refer to the intentional introduction of one dominant error at a predefined type and location within a validated reference solution, while keeping all other conditions as consistent as possible, with pre-identified correction pathways and numerical consequences. Errors may arise in conceptual understanding, model assumptions, computational implementation or evidence interpretation, and are not equivalent to factual hallucinations. Simple tasks adopt a single error; comprehensive tasks allow interactions among multiple errors, with dependency relationships marked in the instructor version.

The teaching sequence first requires students to state their preliminary judgements and confidence levels, before presenting calculation results or counterexamples relevant to their judgements to direct cognitive conflicts toward specific engineering conditions. Students then revise their conclusions through evidence checking and recalculation, and finally, new tasks are used to examine whether transferable inspection rules have been established. Insights from error-based learning feedback<sup>[2]</sup> and cognitive load research<sup>[3]</sup> jointly suggest that task scope should be controlled to avoid introducing unfamiliar software, complex structures and multiple concealed errors simultaneously. What is proposed herein is a hypothesis for the teaching mechanism, whose effectiveness needs to be verified in follow-up studies.

Figure 1. Overall Framework of Controlled-Error-Driven Teaching



**Research Hypothesis:** Evidence verification and reflective revision support independent judgement. The teaching effectiveness remains to be tested using classroom data.

## 2.3 Evidence-Based Adoption and Scaffold Fading

Rejecting all AI recommendations does not equate to improved engineering judgment. Classroom materials include correct, erroneous, and insufficient-evidence items, requiring students to decide whether to adopt, revise, or suspend judgment based on available evidence. Definition checklists and verification procedures are provided in the early stage; only checkpoints are prompted in the intermediate stage, and all prompts are removed in the later stage. Independent extraction and spaced retrieval are arranged after practice to test knowledge retention. Reviews of relevant learning techniques can inform the design of assessment timing<sup>[4]</sup>.

## 3. Generation and Quality Control of Controlled Error Cases

### 3.1 Domain-Specific Error Taxonomy and Case Architecture

The case library is organized along four threads: load, stress, fatigue modelling, and result interpretation. Each case stores task conditions, reference solutions, error locations, error mechanisms, expected consequences, evidence sources, correction notes, and transfer tasks. Full annotations are retained in the instructor version, while error locations and categories are hidden in the student version. The term “correct” within each case is constrained to the given conditions and cannot be generalized as universal conclusions for physical structures. Table 2 lists the core categories to be developed.

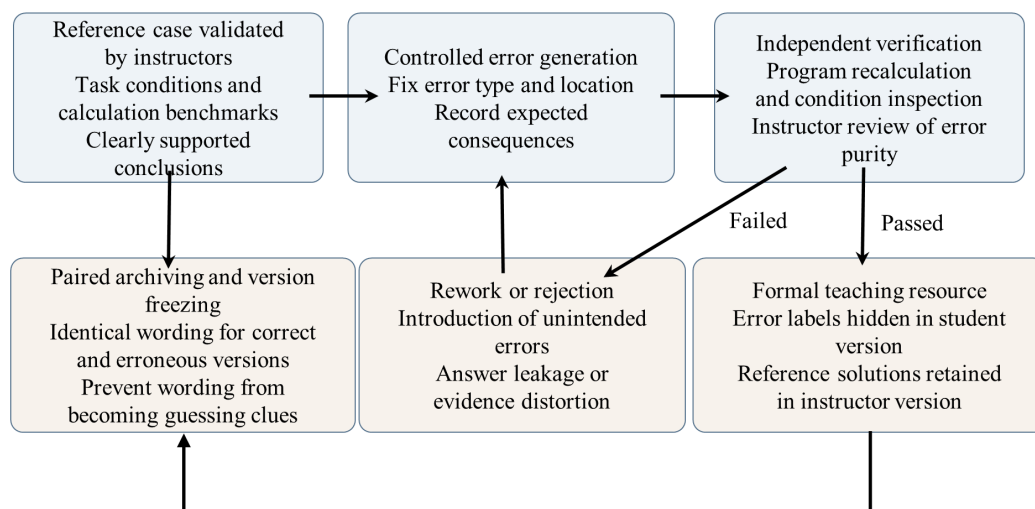
Table 2. Categories of Error Cases for Structural Safety Judgment of Railway Vehicles

| Category                   | Example Error                                                       | Primary Verification Method                               | Learning Objective                               |
|----------------------------|---------------------------------------------------------------------|-----------------------------------------------------------|--------------------------------------------------|
| Conceptual metric mismatch | Substituting stress amplitude into stress-range-based curves        | Check definitions and unify variable metrics              | Matching magnitudes with models                  |
| Curve applicability        | Combining nominal stress with incompatible curves                   | Verify joint classification and stress evaluation methods | Identifying evaluation boundaries                |
| Dynamic approximation      | Using a constant factor to cover the full excitation frequency band | Frequency conversion and response estimation              | Examining model simplifications                  |
| Numerical interpretation   | Directly applying singularity peak values for life calculation      | Inspect mesh and stress extraction schemes                | Distinguishing numerical credibility             |
| Data processing            | Treating the number of sampling points as cycle counts              | Trace time histories and cycle-counting rules             | Maintaining data traceability                    |
| Evidential inference       | Judging fatigue hazard merely from contour plot colours             | Check legends, units and load cases                       | Avoiding visual cues as substitutes for evidence |
| Conclusion extrapolation   | Declaring whole-vehicle safety based on satisfactory local checks   | Examine failure modes and load-case coverage              | Restricting the scope of conclusions             |

### 3.2 Synergy of Retrieval-Augmented Generation and Targeted Mutation

The proposed technical workflow consists of three stages: reference solution construction, controlled mutation, and independent review. First, a retrieval repository is built from instructor-authorized textbooks, lecture notes, and licensed technical documents, with source, version, page number and applicable scope archived. Retrieval-augmented generation links the generation process to external knowledge<sup>[5]</sup>; however, successful retrieval hits alone cannot validate the correctness of conclusions. The model is prohibited from fabricating citations for documents whose versions or clauses cannot be verified. Second, the model is supplied with validated correct cases, target error types, editable fields and fixed fields, to generate erroneous versions with similar linguistic style. For instance, only the input parameter of the fatigue curve is altered from stress range to stress amplitude, while cycle counts, reference curves and calculation methods are frozen. The model also returns change notes visible only to instructors. Elementary cases are constrained to contain merely one dominant error to avoid unintended introduction of unit, symbol or arithmetic mistakes.

Figure 2. Reference-Solution-Driven Workflow for Generation and Review of Controlled Error Cases



Finally, the program recalculates numerical results via independent formula implementation, a rule checker verifies dimensions and input ranges, and instructors audit the engineering semantics and teaching difficulty. Reflective generation can assist the model in revising drafts based on feedback<sup>[6]</sup>. Research on the structural durability of modern equipment generally

adopts integrated simulation–experimental approaches. Su et al. combined CFD simulations with bench tests to reveal the inherent mechanisms of wear suppression and surface restoration enabled by eco-friendly GO nanolubricants, offering an example for the integrated case teaching of "simulation + experiment" in engineering majors [7]. Nevertheless, neither model self-critique nor consensus among multiple models can serve as proof of correctness. Frozen versions are adopted for formal teaching. If supplementary variants are generated during learning activities, they must undergo the same review process or be selected from a pre-vetted variant pool. Figure 2 illustrates the resource development workflow.

### 3.3 Paired Materials and Engineering Authenticity

Correct and erroneous versions shall be comparable in length, term density and figure layout, to eliminate non-technical cues such as “confident wording implies error” or “attached citations guarantee validity”. For initial exercises, instructors may allocate one-third of materials to each category: correct, erroneous, and insufficient evidence. The proportions can be adjusted afterwards with hidden distribution information. Such ratios are pedagogical design choices that should be tuned according to students’ prior knowledge and pilot teaching feedback, and they do not represent real-world AI error rates.

Case review comprises at least four checks: the target error is genuinely present; other conditions remain unchanged without unintended alteration; the reference solution sufficiently supports the target judgement; and correction materials can explain the root cause of the error. The number of generated candidates, approved cases, review time per case and reasons for revision are recorded to evaluate whether AI reduces the total cost of resource development. Engineering source data requires desensitization. Manually defined, analytically computed and experimentally measured data are marked with distinct labels. The three cases presented in this paper are all clearly labelled with data attributes.

## 4. Typical Engineering Cases and Analysis of Error Consequences

### 4.1 Misjudgement of Fatigue Damage Caused by Misuse of Stress Metrics

The fatigue assessment of bogie frame welded joints is adopted as the teaching scenario. A single-slope S–N curve based on stress range is defined, with a reference cycle count of  $2 \times 10^6$  cycles, a reference stress range of 80 MPa and a slope of 3. This curve serves as a teaching model to demonstrate the consequences of metric misuse and does not correspond to a joint class in any specific standard. No knee point, cut-off limit, thickness correction or partial safety factor is introduced in this example, and no acceptance conclusion for real bogie frames can be drawn from it.

$$N_i = 2 \times 10^6 \left( \frac{80}{\Delta \sigma_i} \right)^3, \quad D = \sum_i^{n_i} \quad (1)$$

In the equation, stress range is measured in MPa and cycle count in cycles; D denotes linear cumulative damage. The stress amplitude equals half of the stress range. In the erroneous version, the stress amplitude is directly substituted into the stress-range term in Equation (1), while the curve parameters remain unchanged. This results in an 8-fold increase in predicted life at each load level and the damage being reduced to one-eighth of the correct value. Table 3 presents the full recalculation data.

*Table 3 Teaching Example of Fatigue Damage and Recalculation Results*

| Stress Range/MPa | Cycle Count/ $10^4$ | Correct Life/ $10^4$ | Correct Damage | Damage Obtained by Misusing Amplitude |
|------------------|---------------------|----------------------|----------------|---------------------------------------|
| 60               | 120                 | 474.07               | 0.253125       | 0.031641                              |
| 80               | 80                  | 200.00               | 0.400000       | 0.050000                              |
| 100              | 50                  | 102.40               | 0.488281       | 0.061035                              |
| Total            | 250                 | —                    | 1.141406       | 0.142676                              |

Under the artificially defined criterion of  $D \leq 1$  for this case, the erroneous analysis yields  $D = 0.142676$  and claims compliance, whereas the correct calculation returns  $D = 1.141406$ , which cannot support the pass conclusion. The two results differ by a factor of eight, corresponding to an 87.5% underestimation of damage. If all spectrum levels are repeated uniformly at the same scaling ratio, the allowable repetition factors derived from  $D = 1$  are approximately 7.01 and 0.876, respectively. This discrepancy arises from misalignment of variable definitions rather than changes in material fatigue properties. Real-life life criteria and correction conditions shall be specified separately.

The sequential teaching prompts are: what input quantities are required by the model; what quantities are provided by the data; how inconsistency between the two alters the predicted life; whether available evidence supports the pass conclusion. The transfer task reverses the error direction and asks students to identify overly conservative conclusions caused by substituting stress range into a stress-amplitude-based relationship. This arrangement tests students' understanding of variable matching, instead of merely memorizing that "AI always underestimates damage".

#### 4.2 Misjudgement on the Applicability of Constant Dynamic Load Factor under High-Frequency Excitation

The scenario considers wheel polygon excitation approaching the modal frequency of the housing. With a vehicle speed of 300 km/h, wheel diameter of 0.92 m and polygon order of 20, the excitation frequency is approximately 576.65 Hz. To isolate dynamic concepts, an equivalent linear single-degree-of-freedom (SDOF) model excited by harmonic force is adopted. The natural frequency is set to 580 Hz and the damping ratio to 0.03. Local stress of the same mode is assumed to be proportional to the generalized displacement. This model only demonstrates the applicability limitation of constant coefficients and does not represent the full transmission path of a vehicle.

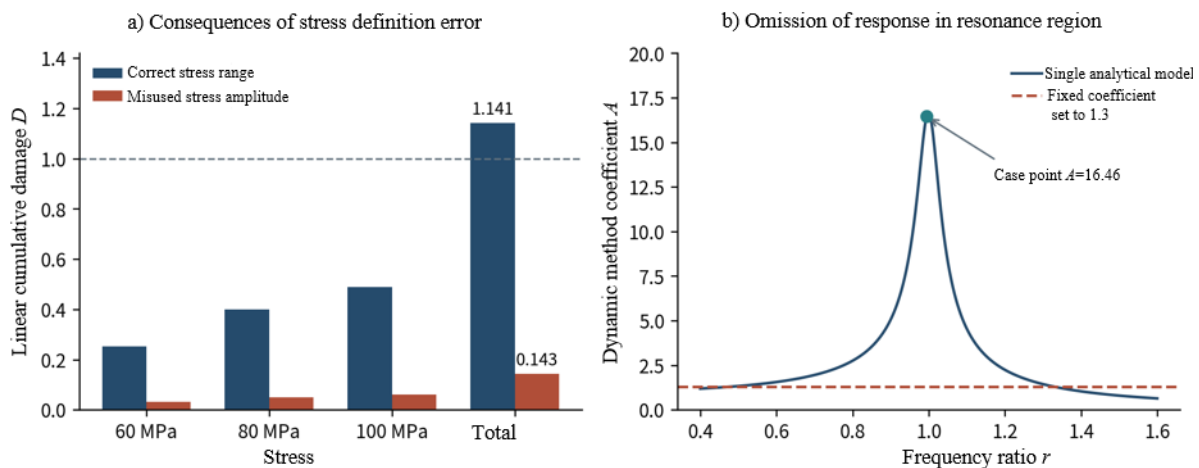
$$f_e = \frac{kv}{\pi d}, \quad r = \frac{f_e}{f_n}, \quad A(r, \zeta) = \frac{1}{\sqrt{(1-r^2)^2 + (2\zeta r)^2}} \quad (2)$$

In the equation,  $v$  is in m/s,  $d$  in m, and  $k$  denotes the polygon order.  $A$  represents the amplification factor relative to static displacement under the same harmonic force amplitude. Equation (2) cannot be unconditionally applied to base excitation, total acceleration response or multi-mode superposition. Nor can one conclude that dangerous stresses will inevitably occur in real structures merely because the frequencies are close.

Table 4. Inputs and Results of the Near-Resonance Teaching Model

| Item                                         | Preset or Calculated Value | Data Nature                               |
|----------------------------------------------|----------------------------|-------------------------------------------|
| Vehicle speed and wheel diameter             | 300 km/h; 0.92 m           | Teaching preset                           |
| Harmonic order and natural frequency         | 20; 580 Hz                 | Teaching preset                           |
| Damping ratio and static stress amplitude    | 0.03; 2.00 MPa             | Teaching preset                           |
| Excitation frequency and frequency ratio     | 576.65 Hz; 0.99422         | Calculated by Equation (2)                |
| Dynamic amplification factor                 | 16.459                     | Calculated by Equation (2)                |
| Dynamic stress amplitude of equivalent model | 32.92 MPa                  | Static stress amplitude multiplied by $A$ |
| Estimated value using constant factor        | 2.60 MPa                   | Artificially adopting a factor of 1.3     |

Figure 3. Comparison of error consequences and frequency response in teaching examples. Data calculated from Equations (1) and (2).



Under the above conditions, the amplification factor is approximately 16.46 and the dynamic stress amplitude is about 32.92 MPa, which is 12.66 times the value obtained using a constant factor of 1.3. Figure 3 illustrates the frequency dependence.

The core assumption for students to revise is the unfounded premise that “a constant factor is sufficient to cover this load case”, rather than directly declaring housing failure based on this simplified result. Subsequent verification shall supplement the actual excitation amplitude, transmission path, modal stress distribution and validation data.

Transfer questions can be constructed for operating conditions far from resonance or with reduced participation of the corresponding mode, requiring students to explain why the same constant factor exhibits different degrees of rationality under varied conditions. A complete engineering interpretation can only be achieved by simultaneously examining excitation, modes and local responses.

### 4.3 Numerical Interpretation Task: Confusion between Local Peaks and Stabilized Responses

The third case trains the judgement of whether an increase in numerical value implies elevated structural risk. To prevent artificial data from being mistaken for finite-element test results, Table 5 uses two explicit functions to generate the teaching sequence. Define the dimensionless mesh parameter  $\eta = h/(1 \text{ mm})$ . The surrogate for local peak stress equals  $150 \times (8/\eta)0.3$ , and the surrogate for stabilized response equals  $96 - 0.08\eta^2$ , both in MPa. The two functions mimic the trend of monotonically rising local peak values and gradual convergence of another response quantity, respectively. They do not represent finite-element calculations of real bogie frames and are not intended to validate any stress extraction method.

Table 5. Artificially Constructed Sequence for Mesh Response Discrimination

| Mesh Size h/mm | Peak Surrogate/MPa | Stabilized Response Surrogate/MPa | Relative Change of Adjacent Stabilized Quantities/% |
|----------------|--------------------|-----------------------------------|-----------------------------------------------------|
| 8              | 150.00             | 90.88                             | —                                                   |
| 4              | 184.67             | 94.72                             | 4.23                                                |
| 2              | 227.36             | 95.68                             | 1.01                                                |
| 1              | 279.91             | 95.92                             | 0.25                                                |

The erroneous material claims that “the maximum stress from the finest mesh is inherently the most accurate and can be directly substituted into fatigue curves”. The given sequence only justifies further inspection of peak characteristics; it cannot confirm that such values represent physical stress or mathematical singularities. Students are required to propose factors for examination such as geometric sharp corners, point loads, constraints or contact edges, check element types, mesh paths and stress definitions, and then select an extraction scheme compatible with the fatigue assessment method. Convergence of the stabilized response merely indicates that this quantity plateaus within the provided sequence, and does not equate to structural safety or a valid fatigue model.

One legitimate endpoint of this task is the judgement of “insufficient evidence”. Students who simply discard the maximum stress and adopt a lower value cannot earn high scores. A high-quality response must specify which quantity to select, the rationale for selection, what evidence remains missing, and how supplementary validation can be performed.

## 5. Instructional Organization and Learning Support

### 5.1 Curriculum Integration and Classroom Arrangement

This teaching module targets senior students who have completed Mechanics of Materials, fundamental vibration and introductory finite element courses. It can be embedded into courses on structural strength of railway vehicles or fatigue and fracture of engineering structures. Four 90-minute teaching sessions are planned, covering stress metric definition, dynamic approximation, numerical interpretation and comprehensive transfer. This arrangement represents a proposed implementation plan and has not yet been delivered. If students lack adequate prior knowledge, basic examples shall be supplemented first and the complexity of errors reduced.

Individual preliminary judgements are archived before classroom discussions to prevent group conclusions from overriding individual learning evidence. Groups may adopt role division: analyst, verifier and interrogator; however, role rotation does not replace independent assessment. After class, students submit brief reflections to explain their reasoning for revising or retaining judgements, and list priority conditions to inspect when encountering similar problems in future.

Table 6. Organisation Plan for a Single 90-Minute Teaching Session

| Stage                             | Time / min | Student Task                                              | Support and Preserved Evidence                |
|-----------------------------------|------------|-----------------------------------------------------------|-----------------------------------------------|
| Knowledge preparation             | 10         | Confirm definitions and essential fundamentals            | Standardized micro-lecture; basic exercises   |
| Independent preliminary judgement | 10         | State judgement, rationale and confidence level           | AI prohibited; initial draft archived         |
| Evidence verification             | 25         | Source checking and independent recalculation             | Layered prompts; verification log             |
| Comparative discussion            | 15         | Compare justifications for alternative interpretations    | Peer interrogation; record of discrepancies   |
| Argument revision                 | 15         | Rewrite conclusions and quantify consequences             | Archive before-after differences and evidence |
| Instructor-led remediation        | 10         | Confirm applicable boundaries and correct interpretations | Distribute vetted reference solutions         |
| Independent exit task             | 5          | Solve a short transfer problem                            | Completed without AI; responses collected     |

## 5.2 Layered Feedback and Error Remediation

Basic feedback provides locating prompts such as “check the definition of input quantities”. Intermediate feedback requires students to state the formula applicability conditions. Advanced tasks only pose probing questions such as “what evidence would be sufficient to alter the conclusion”. AI prompts can be selected according to annotated error types, yet final safety judgements must never be directly provided. Predefined instructor-authored prompts may be used in early stages to guarantee stable scaffolding strategies; the quality of real-time generated prompts shall be separately validated.

Explicit remediation must be completed at the end of each teaching session. Erroneous versions, root causes and correct interpretations are presented side-by-side, and students are required to redo key procedures. Students with weaker foundations may practise correct examples and materials with explicit single errors before tackling cases with hidden-assumption errors. Error cases shall not serve as direct operational engineering instructions, and classroom tasks cannot substitute design review or inspection of real products..

## 6. Teaching Assessment Instruments and Empirical Validation Plan

### 6.1 Research Hypotheses and Comparison Conditions

Three hypotheses are to be tested:

H1: Compared with conventional instruction using only correct worked examples, controlled-error teaching improves evidence-based judgement scores on unfamiliar AI-free tasks.

H2: When missed detections of erroneous suggestions decrease, the false rejection rate of valid suggestions does not rise by a predefined unacceptable margin.

H3: The quality of evidence verification correlates with subsequent transfer performance.

The correlation in H3 shall not be directly interpreted as mediating causation. The primary outcome is defined as the total score on delayed AI-free transfer tasks; other metrics serve as secondary or exploratory outcomes.

Three groups A, B and C are established:

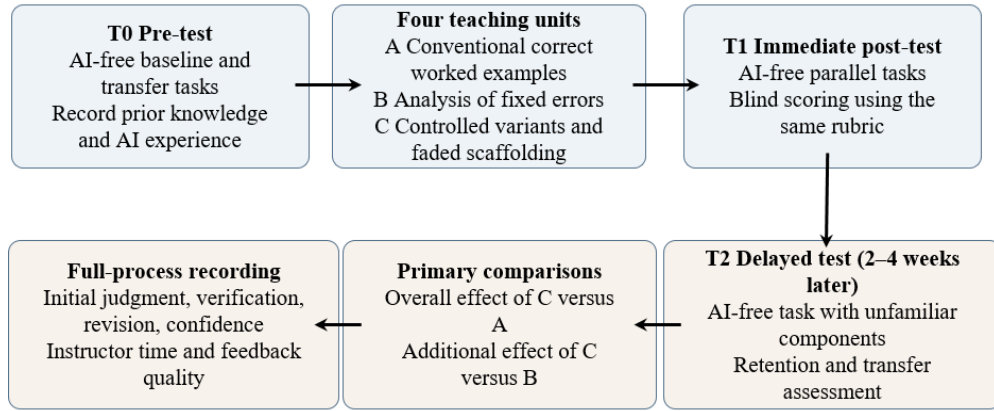
Group A receives conventional correct worked examples and exercises;

Group B uses fixed error cases with standardized instructor feedback;

Group C adopts vetted AI-controlled variants, evidence verification activities and fading prompts.

Total instructional duration, knowledge points, task quantity and instructor contact time are kept identical across groups. Comparison between C and A evaluates the overall teaching package; comparison between C and B assesses the combined value of added variants and support mechanisms. Since Group C modifies multiple factors simultaneously, outcomes cannot be solely attributed to large language models. Factorial studies are required to separate the effects of AI-generated materials and scaffolding strategies.

Figure 4. Proposed teaching comparison and longitudinal evaluation scheme; all arrangements shown in the figure are research design.



## 6.2 Sampling and Measurement Schedule

Instructional conditions are preferentially assigned by intact classes across multiple parallel sections, with stratification based on pre-test results and instructor allocation. Power analysis is performed using the minimal meaningful effect, intraclass correlation and number of classes. Repeated responses from the same student cannot be treated as additional independent samples. If only one class is available per group, the study is framed as exploratory pilot teaching. The confounding between class and treatment effects is explicitly acknowledged, and strong causal conclusions are avoided.

The pre-test T0 records domain fundamentals, performance on analogous tasks and prior AI experience. The immediate post-test T1 examines knowledge mastery. T2, administered 2–4 weeks later, uses unseen components or load cases to assess retention and transfer. AI access is prohibited at all three time points. Parallel test forms with difficulty calibrated through piloting are adopted, alongside consistent formula and reference materials. Optional AI-enabled tasks may be deployed to evaluate human–AI collaboration, but their scores cannot replace the primary outcome. Attrition, missing measurements, cross-group communication and unauthorized extra AI use are all documented.

## 6.3 Engineering Judgement Rubric and Calibration Metrics

Assessment adopts a five-dimensional rubric scored from 0 to 4 points per dimension, with a maximum total of 20 points, which is then multiplied by 5 to convert to a percentage scale. Equal weighting is the initial scheme and requires review and trial scoring by domain course experts. Scoring is grounded on observable student arguments and verification records; a coincidentally correct final conclusion does not compensate for flawed reasoning. Table 7 presents anchor descriptions for scores of 0, 2 and 4. A score of 1 indicates relevant attempts without meeting the level for 2; a score of 3 means most criteria for 4 are satisfied with minor non-critical omissions. When key conditions are misapplied, scoring follows the dimension-specific anchors rather than averaging to compromise.

For sub-items with explicitly labelled correct or incorrect answers, the error detection rate is defined as the number of detected erroneous suggestions divided by the total number of erroneous suggestions. The valid adoption rate equals the count of correctly accepted valid suggestions divided by the total number of valid suggestions, and balanced accuracy is the average of these two metrics. Items with insufficient evidence are reported separately in terms of recognition rate and supplementary verification scores, without forced binary classification. The false rejection rate of valid suggestions is the complement of the valid adoption rate and is reported alongside the missed error detection rate.

$$BA = \frac{1}{2} \left( \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right), \quad BS = \frac{1}{M} \sum_{j=1}^M (p_j - y_j)^2 \quad (3)$$

In Equation (3), TP denotes true positives (correctly detected erroneous suggestions), FN denotes false negatives (undetected erroneous suggestions), TN denotes true negatives (valid suggestions correctly adopted), and FP denotes false positives (valid suggestions falsely rejected). For binary sub-items, students are further asked to assign a probability  $p$  that “the suggestion is correct”. The label  $y=1$  for valid suggestions and  $y=0$  for erroneous ones. BS represents the Brier score; lower values indicate better overall quality of probabilistic judgements. The Brier score is affected by both discrimination and calibration,

so it should not be interpreted purely as a calibration metric. When sample size is adequate, reliability diagrams and mean confidence bias analysis shall be used for supplementary evaluation.

*Table 7. Five-Dimensional Scoring Rubric for Structural Safety Judgment*

| Dimension                | Score 0                              | Score 2                                                          | Score 4                                                                           |
|--------------------------|--------------------------------------|------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| Condition identification | No identification or obvious misuse  | Partial conditions listed with critical items omitted            | Fully identify stress metrics, model and load-case conditions                     |
| Error localization       | Accept or reject merely by intuition | Locate the stage with incomplete explanation                     | Pinpoint specific assumptions or steps and elaborate the error mechanism          |
| Evidence verification    | No auditable basis                   | Available documents or calculations with incomplete verification | Clear sources and completed independent verification or counterexample validation |
| Argument revision        | No revision or new errors introduced | Numerical revision without stating conclusion boundaries         | Quantify consequences and form conditional supportable conclusions                |
| Transfer and reflection  | Mechanically copy original answers   | Able to transfer partial procedures                              | Identify new conditions and propose adapted verification routes                   |

Two raters independently score all submissions or a sufficiently large randomly sampled subset while being blinded to group allocation. Practice scoring using non-study samples is conducted prior to formal assessment. Weighted dimension-wise agreement metrics and the intraclass correlation coefficient for absolute agreement of total scores are reported, together with model specifications and confidence intervals. Pre-consensus scores are retained for reliability analysis, while post-consensus scores are adopted for final evaluation. The rubric still requires expert review, piloting and scoring evidence to establish validity and cannot be claimed as a fully validated instrument directly.

#### 6.4 Statistical Analysis and Implementation Quality

The primary analysis uses T2 transfer scores as the dependent variable, group assignment as the independent variable, and controls for T0 baseline performance, accounting for classroom clustering under the actual allocation structure. Multilevel models may be adopted when the number of classes is sufficient. Adjusted mean differences, 95% confidence intervals and standardized effect sizes are reported, with primary comparisons predefined. Temporal changes across multiple measurement time points can be analysed via models incorporating group-by-time interaction terms as secondary outcomes. A predefined acceptable threshold for the difference in false rejection rates is set for H2; a non-significant difference cannot serve as proof of no adverse impact.

*Table 8. Data Collection and Reporting Requirements for the Empirical Teaching Phase*

| Data Item                            | Collection Timepoint            | Reported Content                                                  | Purpose                            |
|--------------------------------------|---------------------------------|-------------------------------------------------------------------|------------------------------------|
| Sample and missing measurements      | Throughout the study            | Group sizes, number of classes, attrition reasons                 | Comparability assessment           |
| Independent transfer score           | T0 T1 T2                        | Mean, dispersion, adjusted difference and confidence interval     | Primary outcome evaluation         |
| Missed detection and false rejection | T0 T1 T2                        | Classification denominators, proportions and confidence intervals | Evidence-based adoption assessment |
| Probabilistic judgement              | T0 T1 T2                        | Brier score, reliability diagram and confidence bias              | Judgement quality evaluation       |
| Verification and revision            | Each teaching session           | Coded proportions and before-after work differences               | Mechanistic insights               |
| Resources and time                   | Preparation and teaching stages | Review pass rate, total instructor time                           | Implementation cost assessment     |

Process metrics include the proportion of valid verifications, proportion of unsupported acceptances, proportion of substantive revisions, task completion time and instructor review time. Responses to open-ended questions are manually coded into four categories: conceptual correction, evidence supplementation, methodological adjustment and superficial rewriting. Coding definitions are specified and inter-coder consistency tested. For small-sample studies, individual changes and case-based evidence are prioritized; complex mediation models or overreliance on single significance results are avoided. Table 8 lists

data items to be collected in follow-up work; no virtual effect values are filled in this study.

## **7. Merits, Boundaries and Conditions for Scaling**

### **7.1 Sources of Instructional Gains from Generative Artificial Intelligence**

The proposed framework leverages generative artificial intelligence to produce varied expressions and load-case variants for the same error mechanism, making it harder for students to answer merely by memorizing problem wording. Retrieval and tool-calling capabilities support source tracing and recalculation, while layered prompts modulate the intensity of learning support. These functions expand coverage of professional cases. However, instructional gains can only be claimed when material quality after instructor review, preparation efficiency and students' independent performance are improved. In particular, total time spent on AI-assisted resource development should be compared against manual authoring by instructors, rather than only reporting generation speed.

### **7.2 Dual Risks of Error-Based Learning and Automation Dependency**

Erroneous materials may increase cognitive load, and inadequate remediation can leave students with persistent misconceptions. The scaffolding strategy therefore progresses from basic to comprehensive tasks and from single to multiple embedded errors, paired with explicit feedback and delayed retesting. AI-generated rationales may enhance superficial plausibility of answers. Students shall form individual judgements first before verifying recommendations. Automated AI scoring cannot be directly adopted for high-stakes grades. Case difficulty should not be determined solely by model self-evaluation; it shall be adjusted based on expert review, pilot performance and actual completion time.

Excessive scepticism also requires monitoring. If error detection rises while the valid suggestion adoption rate declines, the training may only boost rejection tendency without proven improvement in engineering judgement. Evaluation should focus on balanced accuracy and evidence utilization, while analysing diverse response patterns such as “correct conclusion with flawed reasoning” and “deferred judgement with reasonable supplementary plans”. Students with weak foundations shall not be automatically labelled based on a single low score; instructor review and dynamic adjustment are permitted.

### **7.3 Implementation Limitations and Reproducibility**

The analytical cases presented in this paper illustrate error consequences but exclude complex vehicle models, real material scatter and uncertainties in line service loads. Specifically, the single-mode model and artificial mesh sequence are developed for targeted teaching objectives; their computational clarity must not be mistaken for engineering fidelity. Future work may incorporate authorized and desensitized test or simulation data, keeping inputs, algorithms and reference answers traceable, and gradually introducing multi-source evidence and uncertainty considerations.

Model updates, prompt modifications and retrieval repository revisions may alter case difficulty and feedback style. Model identifiers, generation dates, sampling settings, prompt versions, referenced document editions, audit records and final case files shall be documented, and formal resources frozen during classroom trials. The teaching research undergoes institutional review and participant consent procedures as required. Student identifiers are de-identified, and research participation is separated from normal course assessment. This paper lacks classroom implementation data, rubric validation evidence and cost metrics, which will be supplemented by empirical data in follow-up studies.

## **Conclusion**

First, for railway vehicle structural courses, this paper defines controlled errors as instructional resources built upon clear engineering mechanisms with verifiable reference solutions and remediation pathways. A generation workflow integrating reference-solution derivation, targeted mutation, independent verification and instructor vetting is proposed, enabling AI-generated materials to support professional judgement training.

Second, engineering examples demonstrate that errors in stress definition and dynamic approximation can drastically alter calculation outcomes. In the defined single-slope fatigue model, misuse of stress amplitude leads to an 87.5% underestimation of damage. Under the specified near-resonance single-mode condition, adopting a constant factor omits prominent frequency-dependent amplification. These numerical results demonstrate the professional significance of the pedagogical problem and do not constitute evidence of teaching effectiveness.

Third, an evaluation scheme is established with delayed AI-free transfer performance as the primary outcome, supplemented by error detection, valid adoption, evidence verification and probabilistic judgement. Follow-up research shall examine whether learning gains persist, whether false rejection increases, and whether instructor workload remains acceptable, so as to determine target learners and appropriate support intensity for controlled-error teaching.

## Funding

This work was supported by the Jiangxi Provincial Project for Degree and Postgraduate Education Teaching Reform (YJSJG-2026-05-12)

## Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

## Reference

- [1] Bastani, H., Bastani, O., Sungu, A., et al. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- [2] Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, 68, 465–489. <https://doi.org/10.1146/annurev-psych-010416-044022>
- [3] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. [https://doi.org/10.1207/s15516709cog1202\\_4](https://doi.org/10.1207/s15516709cog1202_4)
- [4] Dunlosky, J., Rawson, K. A., Marsh, E. J., et al. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4–58. <https://doi.org/10.1177/1529100612453266>
- [5] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474).
- [6] Shinn, N., Cassano, F., Berman, E., et al. (2023). Reflexion: Language agents with verbal reinforcement learning. *arXiv*. <https://arxiv.org/abs/2303.11366>
- [7] Su, J., Ouyang, Y. Q., Yin, L. R., et al. (2026). Advancing wind turbine gearbox durability with eco-friendly GO nanolubricants: CFD simulation–experiment synergy for understanding flow dynamics, wear suppression and surface restoration mechanisms. *Tribology International*, 213, 111034. <https://doi.org/10.1016/j.triboint.2025.111034>