Received: 3 May 2025 | Accepted: 24 August 2025

DOI: 10.1111/mice.70063

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING WILEY

RESEARCH ARTICLE

Computers in Education

# Proactive framework for evaluating retrieval-augmented generation-based learning assistants in engineering education

Linxin Hua[1] | Lirui Guo[1] | Nan Zheng[1] | Ye Lu[1] | Jia Xu[2] | Jianghua Deng[3]

[1]Department of Civil and Environmental Engineering, Monash University, Clayton, Victoria, Australia

[2]Department of Information Systems and Business Analytics, Deakin University, Burwood, Victoria, Australia

[3]School of Civil Engineering & Architecture, Changzhou Institute of Technology, Changzhou, Jiangsu, China

**Correspondence**
Nan Zheng, Department of Civil and Environmental Engineering, Monash University, Clayton, VIC 3800, Australia.
Email: nan.zheng@monash.edu

**Abstract**

Retrieval-augmented generation (RAG) enabled learning assistants are promising for engineering education, given their capability to supplement domain-specific knowledge and enhance student support. However, it is also a known problem that RAG demands adequate knowledge bases and can experience unreliable retrieval generation alignment. This study proposes a proactive evaluation framework for RAG-based learning assistants, eliminating the need for student feedback in system evaluation. The framework is demonstrated using a Civil Engineering education tool, CivASK. The evaluation framework identifies the deficiencies in CivASK, including database gap, contextual misunderstanding, and incomplete retrievals, based on the performances under simulated student inquiries, automated retrieval ranking, and expert-validated evaluations. Specifically, 742 student queries are analyzed, and 374 test questions are generated for assessment, showing the practical utility of the proposed evaluation framework for real-world education assist development. The application of the proposed framework is transferable to assist other engineering courses as well.

## 1 | INTRODUCTION

Generative Artificial Intelligence (AI) has demonstrated extensive practical applications across various tasks, particularly in creating new samples from limited datasets to optimize model performance (Lu et al., 2023; Shim, 2024) and understanding complex engineering data (Zhao et al., 2024) to extract critical information (Chun et al., 2023; G. Wang et al., 2023). Recent advancements in large language models (LLMs) have transformed these systems into tools applicable across diverse fields and tasks, enabling unprecedented capabilities in data generation (Guo et al., 2025), analysis (Areerob et al., 2025), and interpretation (Yin et al., 2024; Yong et al., 2023).

Building upon these technological foundations, the educational sector has increasingly adopted AI to develop interactive learning environments (ILEs), where intelligent tutoring systems have emerged as key methodologies for enabling instructors to build comprehensive and supportive learning frameworks. While traditional approaches required manual design of static content structures, LLMs introduce dynamic adaptability through advanced knowledge representation, enabling real-time personalization and responsive content generation. The integration of AI into education has been fundamentally enabled by knowledge engineering (Adeli, 1990a; Paek & Adeli, 1990), encompassing sophisticated knowledge representation (Adeli, 1990b) and reasoning capabilities that simulate human-like understanding and problem-solving (Hua et al., 2024; Karataş et al., 2025). These technologies demonstrate effectiveness across educational domains, from programming instruction (Yilmaz & Karaoglan Yilmaz, 2023) to language learning (Qu & Wu, 2024), through immediate feedback and personalized support mechanisms, with applications continuing to expand across multiple disciplines.

Engineering education has historically leveraged knowledge-based expert system technology to provide structured instructional support, building upon successful applications in engineering design practice (Adeli, 1987). Early implementations, such as expert systems for structural design optimization, demonstrated how predefined knowledge bases could systematically guide complex engineering decision-making processes (Adeli & Balasubramanyam, 1988; Adeli & Hawkins, 1991). These approaches were subsequently adapted to educational contexts, enabling engineering education through interactive systems that provided step-by-step guidance (Waheed, 2000; N. Wang & Adeli, 2012). However, engineering education faces significant barriers due to the lack of reliable, domain-specific knowledge bases for modern AI systems (Akolekar et al., 2025). Borges et al. (2024) demonstrated that Generative Pre-trained Transformer 4 (GPT-4) (Achiam et al., 2024) can answer 65% of university-level science, technology, engineering, and mathematics (STEM) questions, particularly excelling in well-specified problems. Although it is enough to match the average performance of students, it is insufficient to serve as a reliable learning assistant in engineering disciplines. Talha Junaid et al. (2024) identified key limitations in using generative AI for engineering education, particularly insufficient knowledge bases that lack comprehensive engineering specifics, leading to incomplete responses or hallucinations in specialized queries.

To address this limitation, retrieval-augmented generation (RAG) has become a critical approach for integrating domain-specific knowledge into LLMs (Lewis et al., 2020). By retrieving relevant information from external knowledge bases, RAG grounds responses in reliable sources, reducing hallucinations and improving output accuracy. This makes RAG-based assistants promising tools for engineering education (Chen et al., 2024; Salemi & Zamani, 2024). For instance, OwlMentor leverages RAG to assist scientific text comprehension through document-based chats and automated question generation, aligning AI outputs with learning objectives (Thüs et al., 2024). Modran et al. (2025) developed a dataset from laboratory equipment documentation and integrated it with a custom LLM to provide accurate and contextually relevant assistance. These applications show the capability of RAG to mitigate hallucinations and provide factual information when paired with structured knowledge bases, for example, academic textbooks or curated engineering datasets (Areerob et al., 2025; Lloret Pardo et al., 2024). However, knowledge base reliability varies significantly (Chen et al., 2024), with RAG systems often failing when documents become outdated or lack detail, resulting in irrelevant retrievals or incomplete answers (Barnett et al., 2024; F. Wang et al., 2024). Therefore, enhancing knowledge base quality and mitigating low-quality information impact remains crucial (Fan et al., 2025; J. Li et al., 2025).

This dependency becomes especially critical in engineering education, where tools must align precisely with course materials and learning objectives. Database optimization remains challenging as traditional cycles rely on student feedback that is often delayed, general, and insufficient for targeted improvement (Demszky et al., 2024; Röhl et al., 2025). Without actionable insights into query failures, iterative optimization of the knowledge base and retrieval performance becomes inefficient and time-consuming (Lauro et al., 2025; Siriwardhana et al., 2023).

This study proposes a performance evaluation framework for CivASK, a RAG-based learning assistant system (LAS) developed for civil engineering education. This framework comprises three core components: (1) a knowledge base developed from course materials, (2) a question generation module that is based on student inquiries to create test questions, and (3) a validation mechanism assessing RAG-based systems through retrieval quality and response performance. Using this framework, the study analyzes historical student questions and generates course-relevant queries to evaluate the system. A Semantic Coverage Index (SCI) is introduced to assess whether generated test questions comprehensively cover associated notes. The response validation reveals system reliability and typical failure patterns, demonstrating the framework's practical utility for real-world educational assistant applications.

The rest of this paper discusses the development of the LAS (Section 2), the evaluation framework (Section 3), and presents the implementation of the proposed framework and the findings from a practical evaluation of a learning assistant (Section 4), concluding with the contributions and limitations of this study (Section 5).

## 2 | LAS

### 2.1 | Structure of CivASK

In typical Civil Engineering courses, students need to learn and apply complex equations to solve practical problems. This poses significant challenges due to the subtle

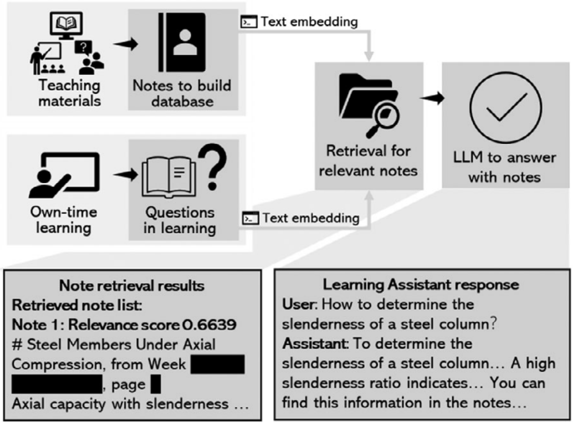


**FIGURE 1** CivASK construction and query workflow with sample note retrieval results and assistant response. LLM, large language model.

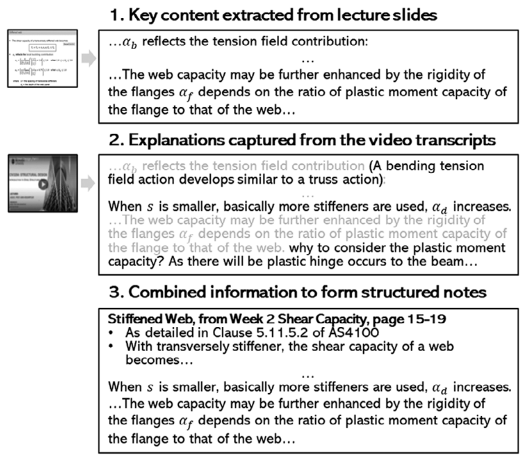


**FIGURE 2** Transformation from lecture slides and video transcripts to structured knowledge notes.

differences in parameter definitions and theories, which are critical for accurate analysis and design (Felder, 1988; N. Wang & Adeli, 2012). In addition, conceptual understanding is essential for connecting theory with practice, for instance, recognizing interactions among structural components, the role of material properties in performance, and the influence of environmental factors on engineering decisions. Without timely and responsive learning, students can easily struggle, affecting their ability to achieve the targeted learning outcomes.

This study presents CivASK, an LAS developed to support students in knowledge retrieval and explanation. The system is built on a RAG-integrated framework, enabling CivASK to retrieve and generate responses based on a course-specific knowledge base. The core methodology of CivASK involves converting the corresponding teaching materials into a structured knowledge base, which is a process potentially applicable for any course (Lewis et al., 2020). For demonstration, a knowledge base was derived from a core Civil Engineering course in Section 4.

Figure 1 illustrates the development and application of CivASK. Teaching materials, including lecture slides, prerecorded videos, and previous semester Frequently Asked Questions (FAQs), are first converted to text and stored in a knowledge base. This knowledge base is then transformed into a vector database using an embedding model (text-embedding-3-small). Designed to support students when teaching staff are unavailable, CivASK converts student questions during self-paced study into vectors using the same embedding model and employs cosine similarity to identify the five most relevant notes. Last, an LLM, gpt-4o-mini-2024-07-18 in the deployment, receives the related notes and the question to generate responses, with temperature set to 0.2 for consistency while allowing controlled variation.

## 2.2 | Knowledge base construction

A typical note includes a title (e.g., "Stiffened Web, from Week 2 Shear Capacity") that specifies the source and location of the content within the course materials, helping students easily navigate to the relevant resources.

Figure 2 illustrates the construction of a note on stiffened web shear capacity, demonstrating the typical transformation from teaching materials to notes. The process begins by extracting key information from concise, list-formatted lecture slides. Detailed explanations from instructor-recorded videos are added in black text, while the original slide content appears in gray. Each note includes a descriptive title indicating its specific location within the course materials.

## 2.3 | Note-enhanced generation

The note in Figure 2, like other notes in CivASK, contains domain-specific knowledge derived from Clause 5.11.5.2 of AS4100-2020 (Standards Australia, 2020) that current state-of-the-art LLMs cannot effectively address without retrieval capabilities. For example, questions such as "How to calculate shear capacity of a web with a transversely stiffener according to AS4100?" receive inadequate responses from gpt-4.1-2025-04-14, as demonstrated below:

> **A**: To calculate the shear capacity of a web with transverse stiffeners according to AS 4100 (Australian Standard for Steel Structures), you need to account for web buckling, shear yielding, and the influence of transverse stiffeners.
> 1. Shear Capacity Without Stiffener (for reference) …
> 2. Shear Capacity With Transverse Stiffeners …
> Step-by-Step Procedure (per AS 4100:2020, Section 6.2.2 and 6.2.3) …

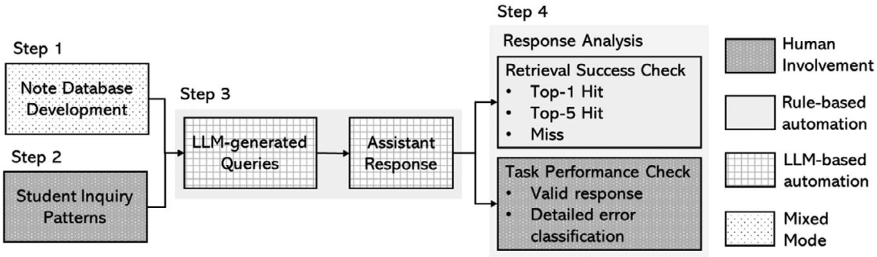


**FIGURE 3** Proactive evaluation framework for retrieval-augmented generation-based learning assistants. LLM, large language model.

In this response, the model incorrectly referenced Sections 6.2.2 and 6.2.3 of AS4100:2020 (Standards Australia, 2020), which contain no relevant content, demonstrating confident hallucination. The response also included irrelevant information such as "Shear Capacity Without Stiffener," failing to address the specific query. However, CivASK (using less advanced model gpt-4o-mini-2024-07-18) avoided such issues through the support of a knowledge database:

> **A**: To calculate the shear capacity of a web with transversely stiffeners, you can refer to the notes titled "Stiffened Web" from Week 2 Shear Capacity, pages 15-19. According to Clause 5.11.5.2 of AS4100, the shear capacity of a web with transversely stiffeners is given by:
>
> $$V_v = V_b = \alpha_v \alpha_d \alpha_f V_w \leq V_w$$
>
> where
> 1. $V_v$ is the shear capacity.
> 2. $\alpha_v$ reflects the local buckling contribution ...
>
> Make sure to also check Table 5.11.5.2 of AS4100 for the relationship between $\alpha_v \cdot \alpha_d$ based on the geometry of the web ...

## 3 | EVALUATION FRAMEWORK FOR LAS

Learning assistant optimization typically relies on user feedback to identify knowledge gaps and usability issues for iterative improvements. However, student feedback is often delayed and lacks technical guidance for improvement. To ensure CivASK delivers timely, aligned, and reliable support, this study proposes a proactive evaluation framework, as Figure 3 details, to systematically identify gaps in retrieval and generation processes.

The framework combines manual and automated processes as labeled in Figure 3. Human involvement tasks are conducted by instructors using their knowledge and experience, including database development, student inquiry pattern extraction, and response validation. LLM-based automation leverages generative capabilities guided by task-specific prompting strategies for question and response generation. Rule-based automation follows user-defined rules for processes such as information retrieval and ranking during the evaluation. Note databases can be developed using mixed approaches, from automated extraction from textbooks to expert-designed structured collections organized by topic (Sajja et al., 2025). The detailed explanations for each key step in the framework are provided below:

Step 1. Note database development: The database comes from course-specific teaching materials converted into structured text chunks as detailed in Figure 2.

Step 2. Student inquiry patterns: Student inquiry patterns are extracted from course materials to guide test question generation, reflecting real-world usage. These patterns may be derived from historical data, such as online discussion forums or in-class student queries, to capture domain-specific question intents. Since different courses have distinct learning focuses leading to varying questioning patterns, these patterns are essential for generating aligned questions. When forum data are unavailable, patterns can be informed by instructor insights or transferred from courses with comparable structures.

Step 3. LLM-generated queries and assistant response: Based on inquiry patterns, the framework generates test questions linked to specific notes in the knowledge base. These questions are submitted to CivASK for note retrieval and response generation.

Step 4. Response analysis: After response generation, the framework involves two validations, retrieval success check and task performance check, which assess the two critical aspects of CivASK: (1) retrieval output quality and (2) task performance of the responses. Retrieval success check evaluates whether the note retrieval identifies the original note from which the test questions derive, reflecting the ability of CivASK to map questions to relevant notes. Retrieval success is categorized as (1) Top-1 Hit, where the original note ranks first in the retrieved list; (2) Top-5 Hit, where the original note appears within the top five results but not the first; and (3) Miss, where the original note is absent from the retrieved list. Task performance check evaluates final responses through manual instructor annotation, comprising: (1) identification of valid responses and (2) detailed error classification based on expert judgment. Given variations in context, model capability, and prompting strategies, manual annotation ensures flexible evaluation grounded in expert analysis.

This structured framework enables performance assessment based on actual requirements of students, facilitating targeted improvements to the knowledge base and response generation capabilities before real-world deployment.

## 4 | EVALUATION OF CIVASK IN PRACTICAL SCENARIOS

This section presents CivASK assessment results under the proposed evaluation framework for a structural design course typically taught in Australian universities. The interdependent nature of the design processes means that a misunderstanding of one concept can lead to accumulated errors in the subsequent learning stages. Without immediate support, students may struggle during self-paced learning, making CivASK critical for providing on-demand query responses. Furthermore, the course includes concepts and parameters that are textually similar yet physically or mathematically distinct, requiring precise understanding to avoid equation misuse. These characteristics make the course an ideal testbed for evaluating CivASK in retrieving contextually relevant notes and generating reliable responses.

With the course selected, historical student queries were first collected and analyzed to identify question patterns. Synthetic questions were then generated based on these patterns and evaluated using the SCI to assess comprehensive coverage of notes. These questions were posed to CivASK for testing, with responses evaluated for both retrieval quality and task performance. Based on identified error types, the study discussed automation implementation of the evaluation process and demonstrated the complete framework through practical application to an additional course.

### 4.1 | Forum inquiry analysis

To ensure the test questions align with student needs, the question patterns were extracted from the online course discussion forum data. The forum is a tool embedded in the online learning system Moodle (Gamage et al., 2022) where students post questions and receive support from teaching staff and other students. A substantial database was compiled for this analysis, comprising 742 student questions collected over six semesters, from Semester 2, 2021, to Semester 1, 2024. All personally identifiable information was removed. Questions were manually classified by topic and intent, in which "topic" refers to the subject matter of the question and "intent" refers to the purpose of the question. For instance, "It mentions on the exam notice that any calculator can be used. Does that mean that online calculators such as … can be used?" addresses the "Administrative" topic with "Clarification" intent. Tables 1 and 2 summarize the classification results, showing categories, definitions, and example questions from the structure design course forum.

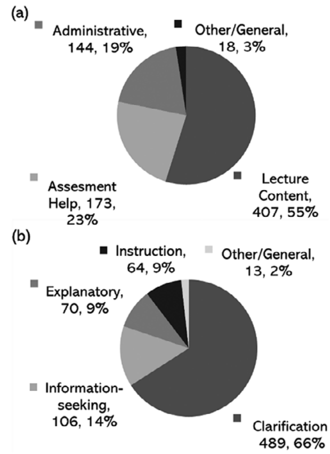


**FIGURE 4** (a) Topic classification distribution and (b) question intent classification distribution over 742 student forum queries.

To quantify question patterns, topic and intent classifications distributions were analyzed. Figure 4a shows 55% of questions focused on "Lecture Content," highlighting the critical role of lecture-related support. "Assessment Help" constituted 23%, while Administrative inquiries accounted for 19%, reflecting practical challenges in assignments and course management. Only 3% fell into "Other/General," demonstrating robust topic classification. This distribution helps instructors understand student needs and guide knowledge base content selection.

Similarly, intent classification in Figure 4b reveals that 66% of questions sought "Clarification," emphasizing the importance of instant validation support in self-paced learning. "Information-seeking" (14%) and "Explanatory" (9%) intents further illustrate the demands for locating resources or understanding design procedures. The minimal "Other/General" intent (2%) suggests the framework effectively captures inquiry intents. Intent analysis helps instructors understand typical question formulation and required assistance types.

**TABLE 1** Topic classification.

| Category | Definition | Example |
|---|---|---|
| Administrative | Questions about course management, policy, or organizational requirements | "It mentions on the exam notice that any calculator can be used. Does that mean that online calculators such as … can be used? Cheers" |
| Lecture Content | Requests to clarify or expand content from lectures/slides | "I am not very sure about the unit in question 2 a) fy is mpa which can be shown as 'n/mm^2' and the unit of t is 'mm'. so we got unit of 'N'. However, this equation is for moment. I am confused." |
| Assessment Help | Queries seeking guidance on assignments and quizzes | "Hi, I'm currently attempting Mock Test 1, and was a bit confused with the calculation for ku." |
| Other/General | Inquiries not fitting the above topics | "Hi, I was wondering if there are any particular chapters of the textbook for the steel component of … (Behaviour and Design of Steel Structures) we should refer to supplement our learning of the unit content. Thanks!" |

**TABLE 2** Question intent classification.

| Category | Definition | Example |
|---|---|---|
| Clarification | Requests to confirm understanding of a concept, parameter, or calculation | "Hi …, could you please clarify what the following sentence means: assume that beams 1 and 2 are sufficiently reinforced in shear (you don't need to consider shear failure for beams 1 and 2) whereas beam 3 has no shear reinforcement (i.e. the failure type to consider is shear). Thanks!" |
| Instruction | Seeking step-by-step guidance for tasks or calculations | "Hi, I am struggling to figure out how to calculate the design action by hand for the timber beam. How will this be done?" |
| Explanatory | Asking for the reasoning behind a method, parameter choice, or principle | "Hi, in Week 3's interactive lecture, for example 2, the lambda ef and ew were … Why do we use these values underlined in red and why not use the same values for both?" |
| Information-seeking | Asking for where to find specific information in course materials | "Hey im having trouble finding the design section capacities for 300plus sections, where can i find them?" |
| Other/General | Inquiries not fitting the above intents | "Please note this is not a question, it is an announcement to the Civil Engineering student community …" |

## 4.2 | Question and response generation

From an educational practice perspective, assessment ("Assessment Help") and course management queries ("Administrative") vary each semester and may involve sensitive information, posing hazards of potential misconduct. Given that most student inquiries centered on "Lecture Content," as shown in Figure 4a, CivASK evaluation focuses exclusively on lecture content. The knowledge base comprises 81 notes covering the lecture slides, prerecorded videos, and FAQs on steel structure design from this course.

A total of 374 test questions were generated by prompting gpt-4o-mini-2024-07-18, as detailed in Table 3, to create questions based on individual notes, as identified as "note chunk" in Table 3, aligned with the question intents from the forum analysis in Section 4.1. The model had no access to the student questions for generations, meaning it did not directly reproduce student queries. It is because real student questions sometimes lack clear and sufficient information, making it hard for LLMs to learn from.

The 374 test questions were posed to CivASK, which retrieved the top five relevant notes, ranked by cosine similarity. Responses were generated using the template prompt, as detailed in Table 4, with an average total response time of 4.84 s, comprising 0.65 s for retrieval and 4.18 s for response generation.

## 4.3 | Question coverage evaluation

LLMs are used to generate test questions based on notes. These questions are supposed to cover all critical information in the notes, but it is challenging in practice. This study proposes the SCI, which assesses whether note sentences are covered by test questions and transfers coverage scores between related sentences within the same note.

**TABLE 3** Prompt template for question generation and examples of typical generated questions.

| Role | Template prompt | Examples of generated questions |
|---|---|---|
| Developer | You are a question-generation system that generates questions based on provided technical information. The generated questions should be in the tone of engineering students who are not familiar with the material. | 1. Clarification: "Can you confirm that I should obtain the yield strength fy and ultimate strength fu from the BHP table for standard sections and from Table 2.1 of AS4100 for built-up sections?"<br>2. Instruction: "Could you provide a step-by-step guide on how to calculate the uniform pressure distribution on the flexible base plate?"<br>3. Explanatory: "Why is it important to consider the critical failure line and the effect of staggered bolts when calculating the net area (An)?"<br>4. Information-seeking: "Where can I find more detailed examples or case studies on calculating modified member slenderness in axial compression?" |
| User | Below is the information of a note: [note chunk]Based on this information, generate several questions that specifically target the note.Each question should be written from the perspective of an engineering student who is learning the teaching material during the own-time learning process.There are four types of questions.<br>1. Clarification: Requesting confirmation that one's interpretation or assumptions regarding a concept or calculation are accurate<br>2. Instruction: Requesting detailed, step-by-step guidance or instructions for design process or calculation<br>3. Explanatory: Requesting an explanation of the reasons behind a calculation method, parameter choice, or underlying conceptual principles<br>4. Information-seeking: Asking for where to find specific information when students do not know where to find the required information | |

**TABLE 4** Prompt template for response generation and an example of question–response pair.

| Role | Template prompt | Example of generated question–response pairs |
|---|---|---|
| Developer | You are a learning assistant for steel design in [course]. When students look at lecture slides or videos, they might be confused or not familiar with the content. They will ask you to provide clarification, guidance, explanation, or related information regarding particular knowledge. When you reply, you must tell the user from which week, which slide, and which page they can find the given information. This info is given in the title of each note. | **Q**: What specific characteristics should a stiffener have to be classified as providing "full lateral restraint?"<br>**A**: To be classified as providing "full lateral restraint," a stiffener should fully restrain the movement of the critical flange. This means that it should effectively prevent both lateral movement and twisting of the flange. A good stiffener that provides full lateral restraint is typically one that is capable of achieving this complete support.<br>You can find this information in Week 1, Section Classification and Bending Capacity, on page 27 of the notes. |
| User | Use the below notes to answer the question:<br>• Question: [student query]<br>• Notes: [retrieved note list]<br>If the answer cannot be found in the notes, reply "I could not find an answer." | |

Each individual sentence from notes is referred to as a "note sentence" for assessment. First, similarity between note sentences and generated test questions is computed using cosine similarity, denoted as $similarity(n_i, q_j)$, where $n_i$ and $q_j$ represent the vector embeddings of note sentences and questions, respectively. When the similarity between a note sentence and a question exceeds a predefined similarity threshold $t$, the question is considered relevant to the note sentence, with the number of relevant questions recorded as $N_i$. The independent SCI for each note sentence is calculated as Equation (1):

$$SCI_i = \max\left(similarity\left(n_i, q_j\right)\right) \times \log\left(N_i + 1\right) \quad (1)$$

The $\log(N_i + 1)$ term prevents the quantity of relevant test questions from disproportionately affecting results. A high SCI score means the note sentence has highly relevant questions as indicated by $\max(similarity(n_i, q_j))$ and multiple relevant questions as indicated by $\log(N_i + 1)$, suggesting comprehensive information examination. A low SCI score indicates the note sentence lacks relevant questions or serves primarily structural or formatting purposes with minimal substantive content warranting question generation.

It is noted that some note sentences may receive low $SCI_i$ scores but maintain strong semantic relationships with other sentences within the same note, indicating indirect coverage when related sentences are covered by test questions. For instance, a structural strength formula may be accompanied by explanatory text, with generated questions targeting explanations rather than formula parameters directly. Since formulas and explanations are highly connected within notes, formula sentences should be considered covered through this indirect relationship despite low direct similarity with test questions.

**TABLE 5** Parameter selection for Semantic Coverage Index.

| Parameters | Description | Values |
|---|---|---|
| $t$ | Note-question similarity threshold | 0.5 |
| $\tau$ | Intra-note similarity threshold | 0.5 |
| $\alpha$ | Propagation strength | 0.6 |
| $k$ | Steepness parameter for sigmoid function | 4 |

To address such issues, intra-note similarities are computed between note sentences to implement score propagation among highly similar sentences. This begins by computing pairwise cosine similarities between all sentence pairs within each note using vector embeddings. For each target sentence, other sentences in the same note are identified as potential propagation sources and ranked by descending similarity for iterative updates. The iterative update mechanism follows a sigmoid-based growth function as Equation (2) shows:

$$SCI_{new} = SCI_{current} + (SCI_{source} - SCI_{current}) \times \sigma\,(similarity - \tau) \times \alpha \tag{2}$$

where $\tau$ represents the similarity threshold parameter between note sentences, and $\alpha$ controls the propagation strength. The sigmoid function serves as a gating mechanism as Equation (3) shows (Ren & Wang, 2023):

$$\sigma\,(x) = \frac{1}{1 + e^{(-kx)}} \tag{3}$$

where $k$ is the steepness parameter. The sigmoid effectively filters out low-similarity propagation attempts while allowing score transfer between highly related sentences. Therefore, the iteration process implements a saturation mechanism where the growth term $(SCI_{source} - SCI_{current})$ diminishes as the score of the target sentence increases.

For the evaluation of generated test questions, the parameters used in this demonstrated case were empirically determined and are presented in Table 5.

The knowledge base of the demonstrated CivASK contains 81 notes, comprising 727 note sentences. Figure 5 shows the distribution of their SCI scores. Manual inspection by skilled instructors identified 57 (7.8%) note sentences lacking relevant questions, classified as untested. Figure 5 demonstrates the proportion of untested note sentences across score ranges, where approximately 80% of sentences with SCI below 0.2 are uncovered by test questions, while untested proportions rapidly decrease as SCI scores increase, approaching near-zero coverage gaps. This indicates that the proposed SCI parameter can effectively identify potentially uncovered portions of the knowledge

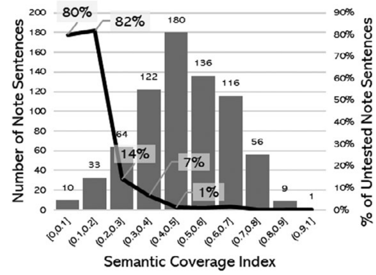


**FIGURE 5** Semantic Coverage Index distribution of note sentences.

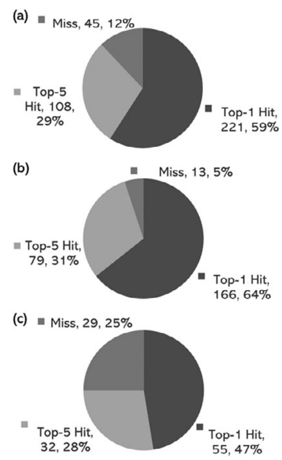


**FIGURE 6** Retrieval success distribution in (a) all responses, (b) successful responses (valid responses identified by task performance check), and (c) failed responses.

base and facilitate targeted manual supplementation of test questions for untested segments.

## 4.4 | Response evaluation

The 374 generated test questions were submitted to CivASK for response generation. The evaluation framework was applied to assess both retrieval accuracy and response validity. A skilled instructor manually evaluated all question-answer pairs by comparing retrieved notes with original content to determine response adequacy and identify error types. Additionally, relevant real student questions from previous semesters were analyzed for comparative validation of the test question coverage.

### 4.4.1 | Assistant response to test questions

Following the evaluation process proposed in Section 3, the results are summarized in Figures 6 and 7. As shown in Figure 6a, CivASK achieved 59% Top-1 Hit and 29% Top-5 Hit retrieval success, demonstrating effective question-note linkage. However, retrieval success alone

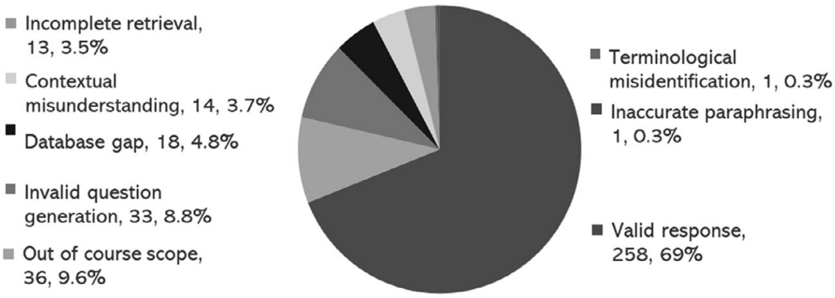


**FIGURE 7** Task performance check, error type distribution of the test questions.

**TABLE 6** Success and error types in task performance check.

| Group | Type | Definition |
|---|---|---|
| Valid response | Valid response | The assistant fully and accurately addresses the query. |
| Question generation issues | Out-of-course scope | The question requests information beyond the scope of the course. |
| | Invalid question generation | The generated question is flawed or nonsensical. |
| System capability limitations | Database gap | The question addresses a valid course topic, but the note database lacks coverage. |
| | Contextual misunderstanding | The assistant misapplies retrieved notes, assuming relevance to an unrelated concept. |
| | Terminological misunderstanding | The assistant misidentifies terms. |
| | Inaccurate paraphrasing | The assistant distorts the content of retrieved notes. |
| | Incomplete retrieval | Critical notes are missing from the top-5 results, leading to incomplete answers. |

inadequately reflects system performance. In Miss cases, the retrieval module does not return the target note; however, the whole list of retrieved notes may still contain relevant information for generating a valid response as a knowledge point may exist across multiple notes in the knowledge database. This disconnect between retrieval and response quality is evident in Figure 6b,c. While 31% of valid responses emerge from Top-5 Hit retrievals and 5% from complete misses (Figure 6b), approximately 75% of invalid responses occur despite successful retrieval of relevant information (Figure 6c). These findings highlight that retrieval accuracy does not guarantee response validity.

The task performance check evaluates how effectively the system generates responses. Table 6 summarizes the three groups along with their corresponding types: successful responses ("valid response") and identified errors, which are categorized into "question generation issues" and "system capability limitations."

Question generation issues include two types of errors arising from question generation module limitations: "out-of-course scope" and "invalid question generation." System capability limitations encompass the remaining error types. "Database gap" results from incomplete database coverage, while "contextual misunderstanding," "terminological misunderstanding," and "inaccurate paraphrasing" result from incorrect understanding of retrieved notes. "Incomplete retrieval" results from insufficient retrieval capability or imperfect retrieval logic. Collectively, these five error types represent capability limitations in specific assistant system modules.

The task performance evaluation results are illustrated in Figure 7. Specifically, 69% of the responses were classified as valid, demonstrating that CivASK is capable of resolving the majority of test questions when retrieval and generation align with course content. Notably, 4.8% of failures occurred due to database gaps, where the questions addressed valid course topics that were not covered in the note database. These failures could potentially be addressed through an updated database that includes all essential notes.

A notable subset of errors (7.8% of total responses) was caused by the limitations in retrieval operation, contextual understanding, or reasoning capabilities of CivASK. These errors included incomplete retrieval (3.5%), contextual misunderstanding (3.7%), terminological misunderstanding (0.3%), and inaccurate paraphrasing (0.3%). Additionally, 18.4% of failures were caused by question generation. Specifically, 9.6% of queries fell outside the scope of the course, while 8.8% were unrepresentative of student queries.

### 4.4.2 | Assistant response to real questions

A subset of real student questions was used to query the assistant system to verify whether generated test questions

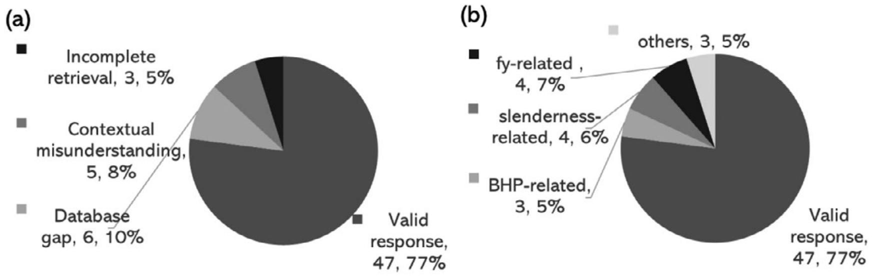


**FIGURE 8** Task performance check of the real questions: (a) error type distribution and (b) knowledge topic distribution of errors.

effectively covered potential student queries and identify database deficiencies or system failures affecting user experience.

Among the 742 real questions, 407 were related to "Lecture Content." Of these, 61 questions covered database content with clear and accurate descriptions, which were submitted to CivASK and manually evaluated by a skilled instructor. Figure 8a shows 77% of the responses ($n = 47$) were identified as valid. Among invalid responses, as shown in Figure 8b, there are 14 invalid question–response pairs, of which 11 pairs involved "BHP table" descriptions, "slenderness calculation details," "fy" acquisition and calculation and the rest three pairs involved other computational details. Error types included database gap (six), contextual misunderstanding (five), and incomplete retrieval (three). Since all questions originated from real students, there is no error related to invalid question generation or out-of-course scope identified. Due to the small sample size, terminological misidentification and inaccurate paraphrasing errors were not observed. Unlike the test questions, which were generated directly from the note database, the student questions collected from real inquiries during past semesters of the selected unit were not explicitly linked to specific notes. As such, a quantitative analysis of retrieval quality was not feasible.

In comparison, after excluding invalid question generation and out-of-course scope categories, the proposed framework identified 47 error responses through the 374 generated test questions. Of these, eight test questions are related to missing "BHP table" descriptions, "slenderness calculation details," and "fy" calculations. These test questions revealed issues that were consistent with those identified from real student questions. This indicates the framework successfully covered the majority (11 out of 14) of the potential improvements identified by real student questions, demonstrating its capability to identify real-world system limitations students would likely encounter.

## 4.5 | Errors in response

Each error category identified in Section 4.4.1 represents a distinct failure type in how CivASK retrieves information and generates answers. These categories diagnose the system's limitations and inform potential refinement areas in content preparation, retrieval accuracy, and response generation. For each identified error type, detailed explanations and illustrative examples demonstrate its characteristics and impacts.

### 4.5.1 | Out-of-course scope

This error type refers to failures that a query exceeds the scope of the course content. In such cases, CivASK is incapable of generating a valid response because the requested information is not covered in the course materials. Such inquiries fall outside the intended scope of the course and thus are not required to augment the database to address them. For example, the following test question asks about "plastic section modulus," which is only mentioned as background knowledge in this unit and not formally covered in the course:

> **Q**: Where can I find more detailed examples or problems related to the plastic section modulus and its application in beam design?

Currently, the system responds to such queries with "I could not find an answer" since no relevant content exists in the database. However, in actual teaching environments, it may be necessary to inform students that the requested information falls outside the course scope. If certain out-of-scope topics are frequently asked by many students, relevant guidance information should be added to the database to facilitate student self-reference.

### 4.5.2 | Invalid question generation

This error type refers to issues where the generated test question is flawed or nonsensical. In these cases, the question fails to represent a realistic or valid inquiry that a student might pose. An example is provided below. This question answers itself by including a reference within the question, eliminating the need for further explanation. Such self-sufficient inquiries are unlikely to represent genuine student queries.

> **Q**: Where can I find Table 5.2 of AS4100 to check the yield limit values for the flange and web?

Another example, as shown below, is a question lacking sufficient context, making it challenging to generate a comprehensive and meaningful response. In this instance, the inquiry references cross-sections A-A and B-B without providing adequate background information, which limits CivASK to provide a coherent and relevant response. This highlights a shortcoming in the current question generation framework, which does not adequately evaluate the clarity or completeness of the conditions presented.

> **Q**: Is it correct that the critical cross-sections A-A and B-B are only relevant when the material yields due to N*?

Collectively, these instances highlight the need to enhance the question generation module, ensuring that the formulated test questions are both logical and sufficiently detailed to elicit meaningful responses.

### 4.5.3 | Database gap

This failure occurs when a question falls within the course scope, yet the note database, which was constructed initially based solely on the teaching materials, lacks certain necessary information to provide a complete answer. In such cases, the teaching staff needs to update and expand the database to cover this missing information. Below is a typical example of a database gap error, presenting the question, part of the retrieved notes, and the response of CivASK:

> **Q**: Where can I find the BHP table to look up the yield and ultimate strengths for my calculations?
> **Retrieved notes**:
> … For standard sections, $f_y$ and $f_u$ are obtained from the BHP table …
> **A**: I could not find an answer.

In this example, the BHP table, which includes geometric parameters of various sections and is critical for assessing structural capacity, is available to students on the course's Moodle site. However, the existing knowledge base contains limited information about the table and lacks sufficient details on i4.5.4ts location and usage, making it inadequate for answering this question. As a result, the retrieved notes contain only a single sentence referencing the BHP table, which is insufficient to support a detailed description, highlighting a clear gap in the database. Once the database is updated to incorporate the relevant information, the inquiry can be effectively resolved:

> **Q**: Where can I find the BHP table to look up the yield and ultimate strengths for my calculations?
> **Retrieved notes**:
> … BHP refers to Broken Hill Pty Co Ltd. You can find BHP table in Week 3, other resources, where we provide a URL for you to download. Or you can check website of Australian Steel Institute for more information (*new note added into the database*) …
> **A**: You can find the BHP table in Week 3, under other resources, where a URL is provided for you to download it. Alternatively, you can check the website of the Australian Steel Institute for more information. The BHP table contains different values of $f_y$ (yielding capacity) for different cross-sections …

Similar challenges can typically be addressed through meticulous augmentation of the note database by the instructor. As evidenced in Figure 8a, which shows the results of testing the assistant with real student questions, database gaps are a significant cause of errors. In educational settings, timely and effective database updates may be essential for enhancing the system's performance and ensuring a positive student experience.

### 4.5.4 | Contextual misunderstanding

This error type addresses issues where the assistant misinterprets or misapplies the information from the retrieved notes, leading to incorrect or irrelevant responses. There are two primary manifestations of this problem: paraphrasing inadequate information and failure to utilize information.

In the first case, the assistant rephrases content from the notes even if that content does not effectively answer the question. For example, as detailed in the following sample, the retrieved notes fail to explain the reasons for using the ratio in rectangular sections. Instead, the assistant simply paraphrases the relevant content from the notes, reflecting an over-reliance on the provided material rather than demonstrating a capacity for deeper analysis. This limitation indicates the system's constrained capability for understanding and reasoning based on the provided content.

> **Q**: Can you explain why the ratio $f^*_{vm}/f^*_{va} = 1.5$ is used for rectangular sections and what it signifies in the context of shear capacity?
> **Retrieved notes**:
> … $f^*_{vm}$= maximum shear stress = $V^*Q_{\max}/It_w$, $f^*_{va}$ = average shear stress = $V^*/d_w t_w$ … For rectangular sections $f^*_{vm}/f^*_{va}$ = 1.5 …
> **A**: … This ratio signifies the relationship between the maximum shear stress and the average shear stress in rectangular sections. In the context of shear capacity, it indicates that for rectangular sections, the maximum shear stress experienced is 1.5 times the average shear stress. This is important for calculating the shear capacity of the section, as it helps in determining how the shear force is distributed across the web of the section …

In the second case, failure to utilize information occurs when the retrieved notes contain valid information, yet the assistant either explicitly states that it cannot answer the question or provides a response that includes unrelated information. This issue is frequently observed in broad and complex inquiries that involve substantial overlap across multiple notes, such as:

> **Q**: Where can I find the definitions of fy and fu in the context of this section capacity calculation?
> **Retrieved notes**:
> … Section Capacity $N_s = k_f A_n f_y$, as detailed in Clause 6.2.1 of AS4100 … where … $A_g$ can be used as $A_n$ if total hole areas $< 100(1 - f_y/0.85f_u)$% … $f_u$ is for fracture failure …
> **A**: I could not find an answer.

These two parameter symbols, $f_y$ and $f_u$, are repeated extensively across the entire retrieved note list. One of these notes explicitly defines both terms in the context of calculating section capacity. However, CivASK failed to effectively utilize this information. From the perspective of a human instructor, the note already contains the information needed to answer the question. But the assistant considers the note did not provide directly relevant information, leading to an invalid response. This reflects a limitation of the LLM to extract effective information when faced with large amounts of information, indicating inadequate intelligence levels.

One potential solution is to switch to a more advanced model with stronger reasoning and comprehension capabilities. However, such adoption often comes with substantially increased costs and still may not guarantee success in all cases. For instance, in the first example of this section, when using the same retrieval results but replacing the model with o3-mini-2025-01-31, one capable reasoning models currently available through the Application Programming Interface (API), the response generated was as follows:

> **A**: The ratio $f^*_{vm}/f^*_{va} = 1.5$ is used for rectangular sections to capture the effect of non-uniform (parabolic) shear stress distribution within the web distribution … It signifies because …

This new response is similar to the previous one obtained using gpt-4o-mini-2024-07-18. While it is factually correct, it fails to provide meaningful reasons, for example, the derivation process and the physical meaning it represents, but only paraphrases the notes. In other words, even more capable models still struggle to recognize when the information provided by the database is insufficient for answering user questions from an engineering education perspective.

Another possible solution, like the approach for addressing database gaps, is to enrich the knowledge base with additional information based on student inquiry needs. For example, in the second case discussed in this section (definition of $f_y$ and $f_u$), a more effective response was obtained after adding more detailed explanations to the relevant notes:

> **Q**: Where can I find the definitions of fy and fu in the context of this section capacity calculation?
> **Retrieved notes**:
> … Section Capacity $N_s = k_f A_n f_y$, as detailed in Clause 6.2.1 of AS4100 … where … $A_g$ can be used as $A_n$ if total hole areas $< 100(1 - f_y/0.85f_u)$% …
> Yield failure—$f_y$—at gross cross-section of this plate.
> Fracture failure—$f_u$—at locations of holes (new added information) …
> **A**: … $f_y$ is the yield strength of the material, which is used in the yield failure context at the gross cross-section of the plate. $f_u$ is the ultimate strength of the material, which is relevant in the fracture failure context at locations of holes.

### 4.5.5 | Terminological misunderstanding

This failure occurs when the assistant incorrectly interprets or fails to associate related terms. In the example below, the test question asks for information on the categories of fasteners. Fasteners connect structural components, and bolts constitute a primary category within the course scope. The retrieved notes encompassed topics such as "Types of Bolts" and "Basic Properties of Bolts," which contain the necessary information relevant to the test question. However, the assistant failed to establish the connection between the term "fastener" and its specific instance "bolt," resulting in an inability to provide an appropriate response.

> **Q**: Where can I find more detailed information on the different types of fasteners used in steel structures and their load transfer capabilities?
> **Retrieved notes**:
> … A steel structure may include: structural members (beams and columns, where the load applied on), connections (where load been transferred), fasteners (where load been transferred) …
> Types of Bolts …
> Basic Properties of **Bolts** …
> **A**: I could not find an **answer**.

To resolve this issue, either employing a more advanced LLM or updating the notes to include additional entries or clarifications that explicitly establish the link between the terms "fastener" and "bolt" would be effective. Below is the response generated using o3-mini-2025-01-31. Notably, the assistant recognized the connection between "fastener" and "bolt":

> **Q**: Where can I find more detailed information on the different types of fasteners used in steel structures and their load transfer capabilities?
> **Retrieved notes**:
> Same **as** before.
> **A**: You can find detailed information on fasteners in the notes on "Types of Bolts" and "Basic Properties of Bolts." In the "Types of Bolts" note, you'll see the common specifications … together with how they are tightened … which directly relates to their load transfer capabilities …

Similarly, it was observed that after adding explicit explanations about the relationship between fasteners and bolts to the database, the assistant was able to successfully retrieve and utilize the correct information. This improvement is evidenced by the following response generated after the database was updated:

> **Q**: Where can I find more detailed information on the different types of fasteners used in steel structures and their load transfer capabilities?
> **Retrieved notes**:
> … Design of Bolts
> Fasteners are hardware devices used to mechanically join or secure two or more objects together. There are numerous types of fasteners, including screws, bolts, nuts, washers, rivets, pins, clips, and anchors, among others …
> **A**: You can find detailed information on the different types of fasteners … Additionally, the load transfer capabilities of bolts can be assessed using specific clauses from AS4100, such as tension capacity (Cl 9.3.2.1) and shear capacity (Cl 9.3.2.2).

### 4.5.6 | Inaccurate paraphrasing

This issue occurs when the assistant correctly retrieves supporting notes and addresses the query but makes errors while restating specific details from the notes. For example, consider the following question and response:

> **Q**: What is the significance of the parameters α_b and α_c in the context of member capacity calculations?
> **Retrieved notes**:
> … Table 6.3.3 (C) provides the values of member slenderness reduction factor ($\alpha_c$) …
> **A**: The parameters $\alpha_b$ and $\alpha_c$ are significant in the context of member capacity calculations as they …
> 1. $\alpha_b$ (Section Constant) …
> 2. $\alpha_c$ (Member Capacity Reduction Factor) …

Although the response adequately addresses the overall requirement by discussing the significance of the parameters, it inaccurately paraphrases the term $\alpha_c$. The correct term, as explicitly stated in the retrieved notes, is "member slenderness reduction factor." This issue does not occur consistently when the same question is resubmitted, suggesting it may result from the inherent stochasticity of the LLM. While the frequency of such errors is low, they remain difficult to eliminate when deploying the system at scale. Recent research suggests that incorporating verification and validation mechanisms can help mitigate such issues, though it also increases response time (Dhuliawala et al., 2024; H. Li et al., 2025). For practical applications, it is essential to find the balance between ensuring response accuracy and a seamless student user experience when implementing such mechanisms and methods.

### 4.5.7 | Incomplete retrieval

This failure occurs when the assistant is unable to find the necessary information to provide an accurate response to a query. In the example below, the assistant failed to retrieve any information that explains the correction factor $k_t$, resulting in a valid response. A review of the retrieved notes reveals that most of the notes contain information on a similar parameter, $k_f$, rather than $k_t$. The only note related to the target parameter lacks any effective parameter explanations itself. This suggests that the retrieval process may have been misled by the textual similarity between $k_f$ and $k_t$.

**Q**: Could you provide a step-by-step guide on how to calculate the effective area using the correction factor kt from the AS4100?
**Retrieved notes**:
… determine $N_t$, $N_t = \min(A_g f_y, 0.85 k_t A_n f_u)$ …
**A**: I could not find an answer.

After directly appending the relevant note, which already existed in the database, to the retrieved notes in this query, the system was able to generate the response below. It is noted that after adding the necessary note, the question can be answered effectively. This indicates that the key to this error lies not in the capability of LLMs but in the accuracy of information retrieval. Such instances are common, as numerous parameters within structural design courses utilize similar symbols and notations. This characteristic of structural design knowledge bases can present substantial challenges for general RAG applications and may require effective retrieval algorithms tailored to the needs of engineering education and the characteristics of course content.

**Q**: Could you provide a step-by-step guide on how to calculate the effective area using the correction factor kt from the AS4100?
**Retrieved notes**:
Eccentricity of connections, $k_t$, from Week 3 Axial Tension, page 13-14.
… For $k_t$ values, check Clause 7.3 of AS4100. Correction factor, $k_t$ is used to … reduce effective area, check Table 7.3.2 of AS4100 …
**A**: To calculate the effective area using the correction factor $k_t$ from AS4100, follow these steps:
1. Determine the Gross Area ($A_g$) …
2. Calculate the Net Area ($A_n$) …
3. Identify the Correction Factor ($k_t$) …
4. Calculate the Tension Capacity ($N_t$) … $N_t = \min(A_g f_y, 0.85 k_t A_n f_u)$ …
5. Determine the Design Tension Force ($N^*$) …

## 4.6 | Automated error identification

### 4.6.1 | Automated error detection in the structural design case study

During academic sessions, instructors can readily apply the proposed evaluation framework to identify database gaps and make targeted updates by adding or revisiting notes before students access LAS. However, the initial deployment of LAS requires comprehensive testing of a large collection of notes, which demands that skilled instructors invest dozens of hours in review and optimization. This substantial time commitment may limit the widespread adoption of LAS and the proposed evaluation framework, particularly in resource-constrained educational settings.

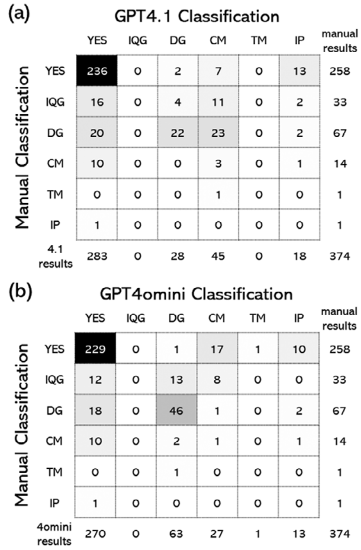


**FIGURE 9** Confusion matrix comparing (a) gpt-4.1 and (b) gpt-4o-mini with manual classification. Error types: invalid question generation (IQG), database gap (DG), contextual misunderstanding (CM), terminological misunderstanding (TM), inaccurate paraphrasing (IP).

To address this challenge, this section introduces an automated error identification module based on error type analysis from Section 4.4.1 as shown in Table 7.

In this module, the LLM is provided with the test question, the list of retrieved notes, and the generated response. Based on this information, the model evaluates the response and classifies any identified errors into one of the predefined error types. The error types "out-of-course scope" and "incomplete retrieval" exhibit similar characteristics to "database gap" (as detailed in Table 6), as all three involve retrieval outputs that lack effective information. Since it is difficult for LLMs to reliably distinguish among them, these error types are consolidated into the unified category of "database gap."

In this section, the module utilized gpt-4.1-2025-04-14 (described by OpenAI as the "smartest model for complex tasks") and gpt-4o-mini-2024-07-18 (described as the "most cost-efficient small model") to evaluate all test questions along with their corresponding retrieval results and responses (OpenAI, 2024). Figure 9 presents the performance of both models in determining whether responses are valid or not, compared to manual evaluation. The results show that both models achieve high accuracy in identifying valid responses, with gpt-4.1 reaching 91.5% and gpt-4o-mini reaching 88.8%. When evaluating overall validity classification (valid vs. invalid), both models

**TABLE 7** Prompt template for error identification.

| Role | Template prompt |
|---|---|
| Developer | You are an expert engineering education evaluator with specialized knowledge in structural design and civil engineering. Your task is to classify responses into exactly one of six categories.You will be provided with 1. A student question, 2. Retrieved notes/information that should be used to answer the question, 3. An AI-generated response.Evaluation process: 1. Analyze the response in the context of the provided question and retrieved notes, 2. Determine if the response is valid or contains errors, 3. If the response is valid, classify it as "Valid response," 4. If the response contains errors, classify it into one of the following error types:<br>• Valid response: The assistant fully and accurately addresses the query.<br>• Invalid question generation: The generated question is flawed or nonsensical.<br>• Database gap: The question is asking for specific information, but the note lacks coverage.<br>• Contextual misunderstanding: The assistant misapplies retrieved notes, assuming relevance to an unrelated concept.<br>• Terminological misunderstanding: The assistant misidentifies terms.<br>• Inaccurate paraphrasing: The assistant distorts the content of retrieved notes. |
| User | Please evaluate the following response for errors:<br>• Question: [student query]<br>• Notes: [retrieved note list]<br>• Response: [response] |

reach approximately 81% accuracy. Given their similar performance but significantly different costs, gpt-4o-mini is recommended.

However, the LLM outputs show significant discrepancies, compared to skilled instructors in determining specific error types. Although both instructors and models evaluate the same items using identical criteria, instructors may employ more sophisticated evaluation processes. Drawing on their cumulative experience, instructors often have their own "standard answers" for most questions, which they use as benchmarks when evaluating the responses. In doing so, they inevitably apply their teaching experience and knowledge beyond the knowledge base and the knowledge from model training.

From the pedagogical perspective, while this automated error identification module has certain limitations in fine-grained error detection, its ability to assess responses can still help instructors reduce manual review workload while maintaining reasonable reliability.

#### 4.6.2 | Automated error identification for conceptual teaching material

To further validate the generalizability of the proposed framework, CivASK was applied to a second case using teaching materials from a Road Engineering unit, specifically incorporating the unit handbook and the first-week introductory materials. The knowledge base was developed following the steps detailed in Section 2 and focuses on unit information, road system classifications and design considerations.

A total of 38 notes were collected from the related teaching materials and embedded in CivASK. Following the question intent classification identified in Section 4.1, the framework generated 125 test questions, which were subsequently submitted to CivASK to generate responses to be evaluated manually and using the automated error identification module proposed in Section 4.6.

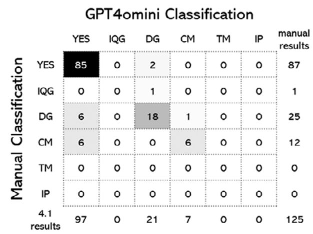


**FIGURE 10** Confusion matrix for gpt-4o-mini and manual classification.

Following Section 4.6.1's recommendation, the analysis uses gpt-4o-mini to check task performance. Figure 10's confusion matrix compares human annotations with automated error detection, showing high agreement in identifying correct responses but weaker performance in classifying error types. For conceptual teaching materials, neither method found "terminological misunderstanding" or "inaccurate paraphrasing," suggesting such errors are rare in this content. This implies LLM comprehension issues may vary by course type, highlighting the need to adapt error classification for different disciplines.

## 5 | CONCLUSION

This study proposes a proactive evaluation framework for RAG-based learning assistants. The framework identifies system errors without student feedback, demonstrated through CivASK, a Civil Engineering learning assistant. The study makes several contributions:

1. This study proposes the first evaluation framework for RAG-based learning assistants in engineering education. The framework uses a knowledge-to-assessment pipeline that transforms educational knowledge bases into evaluation scenarios. Key contributions include: (1) prompt engineering for generating test questions, (2) hybrid validation for assessing retrieval and performance, and (3) identifying system failure modes before deployment. This framework shifts from reactive post-deployment testing to proactive pre-deployment evaluation, enabling systematic assessment without student feedback.
2. The framework was demonstrated through CivASK, a RAG-based learning assistant for Civil Engineering. CivASK converts course materials into structured knowledge bases and integrates lecture slides, video transcripts, and FAQs into a vector database. The system supports self-paced learning with structured, easily navigable content.
3. The proposed evaluation framework was demonstrated through the learning assistant CivASK, which was grounded in a structural design course. The analysis of 742 student forum queries revealed the patterns of student question intent, thereby facilitating the generation of 374 test questions that closely reflect actual student needs. A comprehensive hybrid evaluation uncovered potential limitations and errors in both the retrieval outputs and the final responses. Additionally, validation with 61 real student questions demonstrated that synthetic questions could identify most issues found through authentic student inquiries. The detailed error analysis identified potential ways for system optimization.
4. An automated error identification module utilizing LLMs reduces manual evaluation workload by classifying system responses into predefined error categories. The module achieves high accuracy in identifying valid responses with approximately 81% overall validity classification accuracy. While limited in fine-grained error classification, compared to instructors, it provides a scalable evaluation for resource-constrained educational settings.

While this study advances the evaluation of RAG-based learning assistants, it has limitations requiring future work:

1. The proposed framework tests CivASK with questions from historical data and notes, but real-world use is still unverified. Because generated questions may repeat knowledge base wording, results might be overestimated. Future work should study real student queries in varied situations, such as concept exploration and assignment help. Understanding how students seek support can guide more realistic question creation and system development, extending beyond Q&A to tasks like assessment design and instructional support, while upholding academic integrity.
2. The framework allows customized error categories and metrics, but validation is limited to engineering domains with a fixed knowledge base. This limits cross-disciplinary applicability and statistical analysis. Future work should test the framework across different disciplines and learning systems with larger, diverse datasets to enable comprehensive statistical analysis and establish broader applicability.
3. This work explored automated error identification. The module determines response validity well but poorly classifies specific error types due to complex instructor judgment processes. Future research should explore advanced machine learning and hybrid human–AI workflows to better simulate expert evaluation. Additionally, the framework can guide CivASK optimization through verification mechanisms, enhanced retrieval algorithms, and refined prompt engineering based on identified error patterns.

## ACKNOWLEDGMENTS

Open access publishing facilitated by Monash University, as part of the Wiley - Monash University agreement via the Council of Australian University Librarians.

## REFERENCES

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., … Zhuk, W. (2024). *GPT-4 technical report*. arXiv Preprint arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Adeli, H. (1987). Microcomputer-aided instruction of structural steel design. *Computer-Aided Civil and Infrastructure Engineering*, *2*(1), 75–82. https://doi.org/10.1111/j.1467-8667.1987.tb00246.x

Adeli, H. (1990a). *Knowledge engineering—Volume one—Fundamentals*. McGraw-Hill.

Adeli, H. (1990b). *Knowledge engineering—Volume two—Applications*. McGraw-Hill.

Adeli, H., & Balasubramanyam, K. V. (1988). A novel approach to expert systems for design of large structures. *AI Magazine*, *9*(4), 54–63.

Adeli, H., & Hawkins, D. W. (1991). A hierarchical expert system for design of floors in highrise buildings. *Computers & Structures*, *41*(4), 773–788. https://doi.org/10.1016/0045-7949(91)90187-q

Akolekar, H., Jhamnani, P., Kumar, V., Tailor, V., Pote, A., Meena, A., Kumar, K., Challa, J. S., & Kumar, D. (2025). The role of generative AI tools in shaping mechanical engineering education from an undergraduate perspective. *Scientific Reports*, *15*(1), 9214. https://doi.org/10.1038/s41598-025-93871-z

Areerob, K., Nguyen, V., Li, X., Inadomi, S., Shimada, T., Kanasaki, H., Wang, Z., Suganuma, M., Nagatani, K., Chun, P., & Okatani, T. (2025). Multimodal artificial intelligence approaches using large language models for expert-level landslide image analysis. *Computer-Aided Civil and Infrastructure Engineering*, *40*, 2900–2921. https://doi.org/10.1111/mice.13482

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). Seven failure points when engineering a retrieval augmented generation system. *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering—Software Engineering for AI*, Lisbon, Portugal (pp. 194–199). https://doi.org/10.1145/3644815.3644945

Borges, B., Foroutan, N., Bayazit, D., Sotnikova, A., Montariol, S., Nazaretsky, T., Banaei, M., Sakhaeirad, A., Servant, P., Neshaei, S. P., Frej, J., Romanou, A., Weiss, G., Mamooler, S., Chen, Z., Fan, S., Gao, S., Ismayilzada, M., Paul, D., … EPFL Data Consortium. (2024). Could ChatGPT get an engineering degree? Evaluating higher education vulnerability to AI assistants. *Proceedings of the National Academy of Sciences*, *121*(49), e2414955121. https://doi.org/10.1073/pnas.2414955121

Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), 17754–17762. https://doi.org/10.1609/aaai.v38i16.29728

Chun, P. J., Suzuki, M., & Kato, Y. (2023). Iterative application of generative adversarial networks for improved buried pipe detection from images obtained by ground-penetrating radar. *Computer-Aided Civil and Infrastructure Engineering*, *38*(17), 2472–2490. https://doi.org/10.1111/mice.13070

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2024). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, *46*(3), 483–505. https://doi.org/10.3102/01623737231169270

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024). Chain-of-verification reduces hallucination in large language models. *Findings of the Association for Computational Linguistics ACL 2024*, Bangkok, Thailand (pp. 3563–3578). https://doi.org/10.18653/v1/2024.findings-acl.212

Fan, Y., Liu, L., Sun, N., Wang, Y., Tang, X., & Yu, Z. (2025). *Research on the online update method for retrieval-augmented generation (RAG) model with incremental learning*. No. arXiv:2501.07063. arXiv. https://doi.org/10.48550/arXiv.2501.07063

Felder, R. M. (1988). Learning and teaching styles in engineering education. *Journal of Engineering Education*, *78*(7), 674–681.

Gamage, S. H. P. W., Ayres, J. R., & Behrend, M. B. (2022). A systematic review on trends in using Moodle for teaching and learning. *International Journal of STEM Education*, *9*(1), 6. https://doi.org/10.1186/s40594-021-00323-x

Guo, L., Burke, M. G., & Griggs, W. M. (2025). *Exploring human-SAV interaction using large language models: The impact of psychological ownership and anthropomorphism on user experience*. No. arXiv:2504.16548. arXiv. https://doi.org/10.48550/arXiv.2504.16548

Hua, L., Zheng, N., Lu, Y., Guo, L., & Xu, J. (2024). Use of large language models in engineering education: A case study on infrastructure design report introductions. *35th Australasian Association for Engineering Education Annual Conference*, Christchurch, New Zealand. https://easychair.org/publications/preprint/XpQv/download

Karataş, F., Eriçok, B., & Tanrikulu, L. (2025). Reshaping curriculum adaptation in the age of artificial intelligence: Mapping teachers' AI-driven curriculum adaptation patterns. *British Educational Research Journal*, *51*(1), 154–180. https://doi.org/10.1002/berj.4068

Lauro, Q. R., Shankar, S., Zeighami, S., & Parameswaran, A. (2025). *RAG without the lag: Interactive debugging for retrieval-augmented generation pipelines*. No. arXiv:2504.13587. arXiv. https://doi.org/10.48550/arXiv.2504.13587

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. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada (pp. 9459–9474). https://dl.acm.org/doi/abs/10.5555/3495724.3496517

Li, H., Han, X., Wang, H., Wang, Y., Wang, M., Xing, R., Geng, Y., Zhai, Z., Nakov, P., & Baldwin, T. (2025). Loki: An open-source tool for fact verification. *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, Abu Dhabi, UAE (pp. 28–36). https://aclanthology.org/2025.coling-demos.4/

Li, J., Xu, C., Jia, L., Wang, F., Zhang, C., & Liu, J. (2025). *EACO-RAG: Towards distributed tiered LLM deployment using edge-assisted and collaborative RAG with adaptive knowledge update*. No. arXiv:2410.20299. arXiv. https://doi.org/10.48550/arXiv.2410.20299

Lu, J., Pan, B., Ren, W., Liu, Q., Liu, P., & Oeser, M. (2023). Regeneration of pavement surface textures using M-sigmoid-normalized generative adversarial networks. *Computer-Aided Civil and Infrastructure Engineering*, *38*(16), 2225–2241. https://doi.org/10.1111/mice.12987

Modran, H. A., Bogdan, I. C., Ursuțiu, D., Samoilă, C., & Modran, P. L. (2025). LLM intelligent agent tutoring in higher education courses using a RAG approach. In M. E. Auer & T. Rüütmann (Eds.), *Futureproofing engineering education for global responsibility* (Vol. 1281, pp. 589–599). Springer. https://doi.org/10.1007/978-3-031-83520-9_54

OpenAI. (2024, July). *GPT-4o mini: Advancing cost-efficient intelligence*. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Paek, Y., & Adeli, H. (1990). Structural design language for coupled knowledge-based systems. *Advances in Engineering Software and Workstations*, *12*(4), 154–166.

Lloret Pardo, J., Helsdingen, S. A., & Jermann, P. (2024). Retrieval-augmented generation for finding relevant lectures from quizzes in a multilingual STEM educational environment. *52nd Annual Conference of the European Society for Engineering*, Lausanne, Switzerland. https://doi.org/10.5281/ZENODO.14256811

Qu, K., & Wu, X. (2024). ChatGPT as a CALL tool in language education: A study of hedonic motivation adoption models in English learning environments. *Education and Information Technologies*, *29*(15), 19471–19503. https://doi.org/10.1007/s10639-024-12598-y

Ren, J., & Wang, H. (2023). Calculus and optimization. In J. Ren & H. Wang (Eds.), *Mathematical methods in data science* (pp. 51–89). Elsevier. https://doi.org/10.1016/b978-0-44-318679-0.00009-0

Röhl, S., Bijlsma, H., & Schwichow, M. (2025). Can feedback from students to teachers improve different dimensions of teaching quality in primary and secondary education? A hierarchical meta-analysis. *Educational Assessment, Evaluation and Accountability*, *37*(1), 35–71. https://doi.org/10.1007/s11092-024-09450-9

Sajja, R., Sermet, Y., & Demir, I. (2025). End-to-end deployment of the educational AI hub for personalized learning and engagement: A case study on environmental science education. *IEEE Access*, *13*, 55169–55186. https://doi.org/10.1109/access.2025.3554222

Salemi, A., & Zamani, H. (2024). Evaluating retrieval quality in retrieval-augmented generation. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA (pp. 2395–2400). https://doi.org/10.1145/3626772.3657957

Shim, S. (2024). Self-training approach for crack detection using synthesized crack images based on conditional generative adversarial network. *Computer-Aided Civil and Infrastructure Engineering*, *39*(7), 1019–1041. https://doi.org/10.1111/mice.13119

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2023). Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, *11*, 1–17. https://doi.org/10.1162/tacl_a_00530

Standards Australia. (2020). *Steel structures*. No. AS 4100:2020. https://www.standards.org.au/standards-catalogue/standard-details?designation=as-4100-2020

Talha Junaid, M., Barakat, S., Awad, R., & Anwar, N. (2024). Adopting the power of AI chatbots for enriching students learning in Civil Engineering education: A study on capabilities and limitations. In A. Al-Marzouqi, S. A. Salloum, M. Al-Saidat, A. Aburayya, & B. Gupta (Eds.), *Artificial intelligence in education: The power and dangers of ChatGPT in the classroom* (Vol. 144, pp. 25–47). Springer. https://doi.org/10.1007/978-3-031-52280-2_3

Thüs, D., Malone, S., & Brünken, R. (2024). Exploring generative AI in higher education: A RAG system to enhance student engagement with scientific literature. *Frontiers in Psychology*, *15*, 1474892. https://doi.org/10.3389/fpsyg.2024.1474892

Waheed, A. (2000). A knowledge-based system for evaluation of superload permit applications. *Expert Systems with Applications*, *18*(1), 51–58. https://doi.org/10.1016/s0957-4174(99)00050-0

Wang, F., Wan, X., Sun, R., Chen, J., & Arık, S. Ö. (2024). *Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models*. No. arXiv:2410.07176. arXiv. https://doi.org/10.48550/arXiv.2410.07176

Wang, G., Wang, K. C. P., & Yang, G. (2023). Reconstruction of sub-mm 3D pavement images using recursive generative adversarial network for faster texture measurement. *Computer-Aided Civil and Infrastructure Engineering*, *38*(16), 2206–2224. https://doi.org/10.1111/mice.13020

Wang, N., & Adeli, H. (2012). Web-based tutor for interactive design of single-span and continuous steel beams. *Computer Applications in Engineering Education*, *20*(2), 383–389. https://doi.org/10.1002/cae.20474

Yilmaz, R., & Karaoglan Yilmaz, F. G. (2023). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence*, *4*, 100147. https://doi.org/10.1016/j.caeai.2023.100147

Yin, M., Tang, L., Webster, C., Yi, X., Ying, H., & Wen, Y. (2024). A deep natural language processing-based method for ontology learning of project-specific properties from building information models. *Computer-Aided Civil and Infrastructure Engineering*, *39*(1), 20–45. https://doi.org/10.1111/mice.13013

Yong, G., Jeon, K., Gil, D., & Lee, G. (2023). Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. *Computer-Aided Civil and Infrastructure Engineering*, *38*(11), 1536–1554. https://doi.org/10.1111/mice.12954

Zhao, L., Song, D., Chen, W., & Kang, Q. (2024). Coloring and fusing architectural sketches by combining a Y-shaped generative adversarial network and a denoising diffusion implicit model. *Computer-Aided Civil and Infrastructure Engineering*, *39*(7), 1003–1018. https://doi.org/10.1111/mice.13116

**How to cite this article:** Hua, L., Guo, L., Zheng, N., Lu, Y., Xu, J., & Deng, J. (2025). Proactive framework for evaluating retrieval-augmented generation-based learning assistants in engineering education. *Computer-Aided Civil and Infrastructure Engineering*, *40*, 4651–4668. https://doi.org/10.1111/mice.70063