Discover Computing

Research

Towards an AI tutor for undergraduate geotechnical engineering: a comparative study of evaluating the efficiency of large language model application programming interfaces

Amir Tophel[1] · Liuxin Chen[1] · Umidu Hettiyadura[2] · Jayantha Kodikara[1]

Received: 9 January 2025 / Accepted: 28 April 2025
Published online: 11 May 2025

Abstract
This study investigates the efficiency of large language model (LLM) application programming interfaces (APIs)—specifically GPT-4 and Llama-3—as AI tutors for undergraduate Geotechnical Engineering education. As educational needs in specialised fields like Geotechnical Engineering become increasingly complex, innovative teaching tools that provide personalised learning experiences are essential. Unlike previous studies on AI-driven education, our research uniquely focuses on assessing the role of retrieval-augmented generation (RAG) in improving the accuracy of LLM-generated solutions to Geotechnical problems. A dataset of 391 questions from the related textbook written by Das and Sobhan (Das B, Sobhan K. Principles of Geotechnical engineering, Eight Edition. In: Cengage Learning. 2014) was used for evaluation, with solutions sourced from the textbook's manual. Performance benchmarking focused on 20 challenging questions previously identified by Chen et al. (Chen et al. in Geotechnics 4:470–498, 2024) as problematic for GPT-4 in Zero Shot tasks. GPT-4 with API support demonstrated superior accuracy, achieving accuracy rates of 95% at a temperature setting of 0.1, 82.5% at 0.5, and 60% at 1. In comparison, Llama-3 achieved an accuracy of 25% in Zero Shot tasks and 45% with API support at a temperature setting of 0.1. The findings highlight GPT-4's potential as an AI tutor for Geotechnical Engineering education while demonstrating the need for domain-specific optimisation and advanced formula integration techniques. This study contributes to the ongoing discourse on AI in education by providing empirical evidence supporting the deployment of LLMs as personalised, adaptive teaching aids in engineering disciplines. Future work should explore optimised formula integration strategies, expanded domain knowledge bases, and long-term student learning outcomes.

Keywords Large language models (LLMs) · Application programming interfaces (APIs) · Retrieval-augmented generation (RAG) · Geotechnical engineering education · AI tutors

✉ Liuxin Chen, Liuxin.Chen@monash.edu; Amir Tophel, amir.tophel@monash.edu; Umidu Hettiyadura, ufer0001@student.monash.edu; Jayantha Kodikara, Jayantha.Kodikara@monash.edu | [1]ARC Industrial Transformation Research Hub (ITRH)–SPARC Hub, Department of Civil Engineering, Monash University, Clayton, VIC 3800, Australia. [2]Department of Civil Engineering, Monash University, Clayton, VIC 3800, Australia.

1 Introduction

Artificial Intelligence (AI), particularly through Large Language Models (LLMs) such as GPT-4 and Llama-3, rapidly transforms education by enhancing student engagement, problem-solving capabilities, and personalised learning experiences. LLMs have shown exceptional proficiency in natural language processing and generating structured explanations, positioning them as promising tools for addressing educational challenges across various disciplines [1–6].

Recent studies have assessed LLM performance in STEM-related standardised tests, including engineering licensure exams, graduate entrance exams, and mathematics assessments. These studies offer valuable insights into LLMs' strengths and limitations in structured problem-solving and conceptual reasoning. For example, Pursnani et al. [7] evaluated ChatGPT's performance on the US Fundamentals of Engineering (FE) Exam, showing that prompt modifications significantly enhance AI accuracy, but domain-specific challenges remain. VarastehNezhad et al. [8] assessed GPT-4, Claude 3.5, and Llama-3 on computer science graduate entrance exams, finding that LLMs excel at structured problems but struggle with higher-order reasoning. Similarly, Maitland et al. [9] tested GPT-4 on the MRCP medical exams, revealing high accuracy in structured multiple-choice formats but frequent factual and omission errors in open-ended questions. Additionally, Honig et al. [10] examined student adoption of GPT-powered AI tutors in engineering education, emphasising the need for adaptive AI-human hybrid models to enhance learning engagement. These insights highlight the importance of exploring domain-specific adaptations to optimise the educational utility of AI tools.

Within Geotechnical Engineering, a discipline characterised by specialised knowledge and complex analytical problem-solving, LLMs offer considerable potential. For instance, Xu et al. [11] introduced GeoLLM, a specialised model designed for intelligent geotechnical design automation. GeoLLM has demonstrated the ability to perform calculations related to bearing capacity and settlement analysis, showcasing its effectiveness in processing professional geotechnical knowledge. Similarly, Wu et al. [12] investigated LLM integration into geotechnical workflows and found that these models enhance problem-solving efficiency by supporting decision-making and data interpretation. However, Kumar [3] identified key limitations of ChatGPT in handling geotechnical queries, including hallucinations, unit conversion errors, and challenges in interpreting engineering constraints, suggesting a need for continued refinement of AI tools within this specialised field.

Despite significant advancements, the utilisation of LLMs as educational tools within Geotechnical Engineering remains relatively underexplored. The discipline's reliance on complex formulas and rigorous analytical problem-solving necessitates continuous student guidance, which traditional teaching methods often fail to provide comprehensively [13]. LLMs have the potential to bridge this gap by offering personalised learning experiences and dynamically adapting content to individual student needs [14]. However, implementing AI tutors in geotechnical education presents several challenges:

(1) Institutional Resistance: Universities and educators express concerns regarding student over-reliance on AI and potential academic integrity issues.
(2) Accuracy and Trust Issues: LLMs, including GPT-4 and Llama-3, sometimes generate incorrect responses (hallucinations), particularly in complex problem-solving scenarios requiring engineering judgment [3].
(3) Integration with Existing Learning Platforms: AI tutors must be compatible with learning management systems (LMS), grading rubrics, and curricula, often requiring significant customisation [12].

To enhance the reliability of AI tutors, recent research has explored strategies for improving LLM performance in educational settings. Studies have shown that techniques such as Chain of Thought (CoT) prompting and custom instructional frameworks can significantly enhance model accuracy [7, 15, 16]. The authors' previous study showed how the efficiency of GPT-4 can be improved by incorporating CoT and our proposed custom instruction techniques. The authors showed that by utilising these two techniques, the accuracy of GPT-4 was improved from 29 to 34% and 67% for CoT and custom instruction, respectively [17]. Despite these improvements, other development methods may offer unique benefits that enhance the effectiveness and reliability of AI-driven educational tools. For instance, building a domain-specific LLM from scratch allows for deep customisation but requires extensive resources and expertise in AI training processes, which can be a major drawback [18]. In contrast, fine-tuning existing LLMs on specialised datasets can improve performance in specific domains but risks overfitting and limits broader applicability [19].

Addressing these challenges requires advanced AI-human hybrid learning models and real-time retrieval-based AI frameworks, such as Retrieval-Augmented Generation (RAG), to enhance reliability in AI tutoring [20–23]. This approach balances enhancing response accuracy and specificity through dynamic information retrieval, ensuring that content remains current and aligned with the latest developments in the field [24, 24]. These various development methods, each with their distinct advantages and challenges, are systematically presented in Table 1. This table provides a comparative analysis of LLM development methods, highlighting aspects such as customisation, resource requirements, and scalability.

There are more benefits of using APIs with RAG as follows. APIs with RAG significantly improve the reliability and trustworthiness of AI-generated content. RAG-enabled APIs can cite sources and provide references for retrieved information, akin to footnotes in academic papers. This transparency allows users to verify the information, fostering greater trust in the AI system. Conversely, a standalone LLM, such as GPT-4, operates as a black box with responses that do not inherently include source citations or verifiable references.

Moreover, the flexibility and adaptability of APIs with RAG are advantageous. These APIs allow for hot-swapping various knowledge bases tailored to different subjects or domains, making them versatile tools in education. Educators can integrate specialised databases, technical manuals, or recent research publications to customise the AI's knowledge base to their teaching needs. This adaptability is not feasible with a standalone LLM, constrained by its fixed training data and cannot integrate new knowledge dynamically [14, 25].

In addition, APIs with RAG offer a cost-effective and efficient solution for maintaining and updating AI systems. Retraining an LLM like GPT-4 with new data can be resource-intensive and time-consuming. In contrast, RAG-enabled APIs continuously update their vector databases with new information, ensuring the AI remains current without extensive retraining [26, 27].

This study builds upon the aforementioned advancements by integrating RAG with LLM-based tutoring in Geotechnical Engineering. Key innovations of our approach include:

(1) Context-Aware Problem-Solving: Unlike conventional AI tutors, the RAG-enhanced system retrieves real-time geotechnical references and equations, minimising outdated or inaccurate responses [28].
(2) Improved Conceptual Explanations: AI-generated answers are structured to align with Bloom's Taxonomy proposed by Krathwohl [29], ensuring a step-by-step approach to learning.
(3) Adaptive Learning Responses: The system dynamically adjusts its explanations based on student input, making interactions more effective than static AI tutors.

In addition, we evaluate the efficiency of various LLM APIs, including GPT-4 and Llama-3, as AI tutors in undergraduate Geotechnical Engineering education. We compare these models based on their ability to integrate and apply Geotechnical Engineering formulas, provide accurate explanations, and adapt to different learning styles and needs [30]. Although this research is focused on undergraduate-level problems in Geotechnical Engineering, it offers insights that could potentially enhance understanding of LLMs' applications across broader educational contexts [31]. By providing detailed examinations of these specific areas, the study aims to contribute valuable perspectives on the potential and limitations of LLMs as effective educational tools in the engineering field.

2 Methodology

2.1 Evaluation criteria

The efficiency of the LLM APIs equipped with RAG as AI tutors is evaluated using a set of criteria. These criteria are designed to assess various aspects of the LLMs' performance, pivotal for their effective deployment in educational settings, specifically in Geotechnical Engineering. The evaluation metrics include:

(1) Accuracy of Content: The precision with which the LLMs deliver correct information. This metric is important for the educational AI system because inaccurate content can mislead learners and erode trust in the tool. Prior studies on AI tutors have therefore used answer correctness as a primary performance indicator. For example, Rosoł et al. [32] evaluated GPT-4's tutoring abilities by its accuracy on medical exam questions—reporting about 80% of answers correct. Such findings highlight that high content accuracy is indispensable for effective learning support.
(2) Formula Integration Capability: The ability of LLMs to correctly incorporate and apply Geotechnical Engineering formulas within their responses. In a technical domain like geotechnics, correctly using domain-specific formulas

Table 1 Comparison of three primary LLM development approaches—building from scratch, fine-tuning, and using RAG with API keys

Aspect/Method	LLM development methods		
	Building from Scratch	Fine-Tuning Existing LLM	Using RAG with API keys
Customisation	High: Is fully tailored for specific needs	Moderate: Enhances base models with specific features	Low to Moderate: Varies with API capabilities
Data privacy	Full Control: Complete autonomy over data management	Good Control: Depends on model constraints	Dependent: Subject to third-party policies
Model uniqueness	Tailored: Specifically designed for precise requirements	Beneficial: Utilises a strong base model with added specificity	Standard: Access to the latest models with minimal effort
Resource requirement	High: Requires significant computational power, data, and expertise	Moderate: Utilises existing models and infrastructure	Low: Mainly financial costs for API usage
Time to deployment	Long: Involves extensive data collection and training	Moderate: Reasonably-fast deployment using pre-built models, still requires setup	Short: Quick setup and integration
Technical expertise	Very High: Needs deep ML and NLP knowledge	High: Requires knowledge of model fine-tuning	Moderate: Primarily involves API integration and management
Scalability	Variable: Highly dependent on organisational infrastructure	Variable: Depends on existing infrastructure capabilities	High: Managed by the service provider, easily scalable
Ongoing maintenance	High: Continuous updates and infrastructure maintenance required	Moderate: Needs periodic updates and retraining	Low: Maintenance predominantly handled by the service provider
Costs	Very High: Significant investment in infrastructure and personnel	Moderate to High: Low initial costs but ongoing operational expenses	Cost-Effective: Operates on a pay-as-you-go or subscription basis

Where ML represents Machine Learning and NLP denotes Natural Language Processing is essential for solving problems, so the AI tutor must demonstrate this skill. Previous research in STEM education has explicitly assessed LLMs on their capacity to handle mathematical and engineering formulas, highlighting the importance of this criterion. For instance, Pursnani et al. [7] noted that improvements in ChatGPT's mathematical problem-solving (e.g., applying formulas correctly) were key to tackling engineering exam questions.

(3) Clarity and Utility of Explanations: How effectively the LLMs understandably explain complex concepts and solutions. Evaluating explanation clarity is significant in educational contexts, as a correct answer has limited value if the student cannot follow the reasoning Chiasson et al. [33].
(4) Adaptability to Various Problem Types: The flexibility of the LLMs in handling a diverse range of question types and problem scenarios within the discipline. An effective AI tutor should perform well across different kinds of problems—from straightforward factual queries to complex, multi-step analytical questions. This adaptability reflects the model's generalisation capability, which has been a focus of prior evaluations of AI in education. Studies suggest that AI tutors need to adjust to diverse learning tasks and styles to be broadly useful [34].

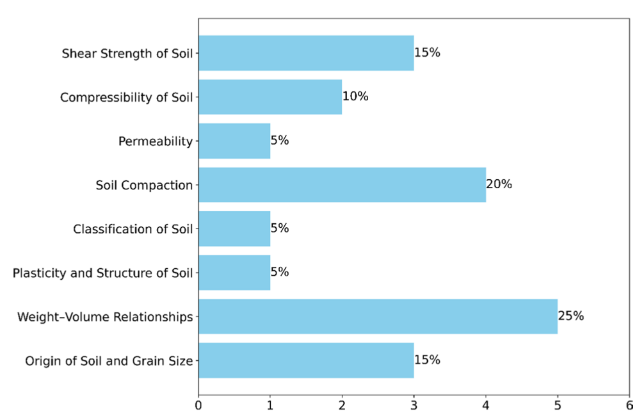
A dataset of 391 questions, sourced from the geotechnical textbook by Das and Sobhan [35], a widely adopted undergraduate textbook, was used as the foundation for evaluation. Solutions to these questions were obtained from the textbook's manual to ensure accuracy. The selection aimed to provide comprehensive coverage of key geotechnical topics, including phase relationships, soil classification, consolidation, and shear strength, ensuring a balanced representation of theoretical and applied problem-solving. Questions were chosen to reflect a range of difficulty levels (i.e., basic, intermediate and advanced), following Bloom's taxonomy, to evaluate AI tutors' ability to recall, apply, and analyse concepts. Furthermore, the dataset includes a variety of calculation-based and conceptual questions to assess AI models' proficiency in formula application, numerical accuracy, and explanatory depth.

Performance benchmarking focused on a subset of 20 challenging questions, previously identified by Chen et al. [17] as problematic for GPT-4, developed by OpenAI [36], under the Zero Shot prompting strategy. These questions were intentionally selected to cover a range of Geotechnical Engineering topics, from phase relationships to shear strength parameters, as shown in Fig. 1. The selected questions—2.3b, 2.8c, 2.9d, 3.5c, 3.6a, 3.7d, 3.7e, 3.12c, 4.5, 5.2, 6.6b, 6.7a, 6.10a, 6.10b, 7.9, 11.4a, 11.18a, 12.8, 12.10, and 12.16—correspond directly to those listed in the textbook. Upon detailed analysis, we found that the majority of questions fell within the lower cognitive levels of Bloom's framework—specifically, Apply (80%) and Analyse (20%), as they primarily required recall and direct application of formulas rather than complex synthesis or evaluation.

We further analysed the semantic structure of the questions:

(1) Symbolic Logic with High Semantic Density: Some questions explicitly specified the variables given and required students to recall a specific equation and perform direct calculations. These questions primarily fell under the Apply level of Bloom's Taxonomy.
(2) Language-Based Logic with Lower Semantic Density: Other questions required interpretation of problem statements, identifying relevant information and determining the appropriate formula or method before performing calculations. These questions required a higher level of conceptual understanding, classifying them under Analyse levels.

Fig. 1 Distribution of questions tested in this study with their percentages across different topics



For this study, all 20 evaluated questions were manually typed into the LLMs as text-based inputs. No questions included graphs, images, or visual representations, meaning the AI models were assessed purely on their ability to interpret and solve textual problem descriptions.

Figure 2 illustrates the distribution of error types in the responses provided by GPT-4 when using the Zero Shot prompting strategy. The chart is divided into three segments, each representing a different type of error:

(1) *Grounding Errors* (60%): This is the largest segment. Grounding Errors indicate that the majority of errors made by GPT-4 under Zero Shot conditions are due to grounding issues, where the model retrieves incorrect equations or constraints for the question posed.
(2) *Conceptual Errors* (20%): This segment represents errors arising from GPT-4's inability to obtain necessary concepts or facts for solving the question.
(3) *Calculation Errors* (20%): This segment, also comprising 20% of the errors, involves mistakes in algebraic and arithmetic operations. Calculation Errors highlight difficulties in the computational aspects of problem-solving by the model.

This visual representation effectively communicates the relative frequency of each error type, underscoring grounding as the predominant challenge faced by GPT-4 in Zero Shot scenarios.

2.2 Overview of APIs

2.2.1 GPT-4 vs Llama-3

When comparing the GPT-4 API and the Llama-3 API, several key differences and advantages emerge, reflecting the strengths and weaknesses of each LLM in practical applications. The GPT-4 API, developed by OpenAI [37], is renowned for its advanced natural language understanding and generation capabilities, built upon extensive training on diverse datasets. This enables GPT-4 to excel in generating coherent, contextually appropriate, and highly accurate responses across a wide range of topics. Its sophisticated architecture supports complex problem-solving and detailed explanations, making it particularly effective for applications in education, technical writing, and customer support. Furthermore, GPT-4's robust API infrastructure provides reliable performance and seamless integration with various software applications, enhancing its usability for developers and end-users. However, it is essential to note that the GPT-4 API is a paid service, which can be a significant consideration for organisations managing their operational budgets.

In contrast, the Llama-3 API with 70 billion parameters, developed by Meta AI, offers a compelling alternative emphasising efficiency and customisation. Llama-3 is free, which can be a significant advantage for organisations looking to implement AI solutions without incurring subscription costs. Additionally, Llama-3 supports fine-tuning, allowing users to adapt the model to specific domains or tasks, enhancing its relevance and accuracy for specialised applications. However, running Llama-3 locally can incur initial setup costs, requiring significant hardware capabilities to operate effectively. This can involve investment in high-performance computing resources, which might offset the cost benefits of the free API. Unlike GPT-4, which runs on the cloud and does not require local hardware investments, Llama-3's local deployment can be resource-intensive. GPT-3.5 was excluded due to its limited performance and the study's focus on evaluating Llama as a free and locally deployable alternative to GPT-4.

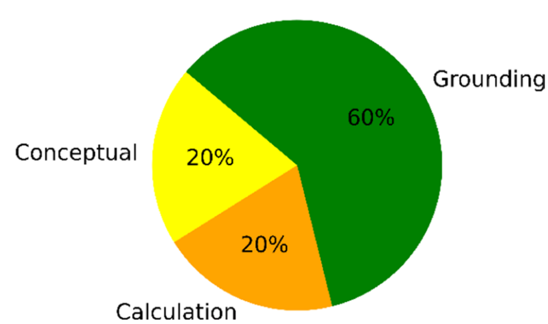


Fig. 2 Error type distribution visualisation in the responses provided by GPT4 using the Zero Shot prompting strategy

2.3 Integration of APIs and RAG using LangChain

LangChain is an advanced framework designed to facilitate the integration of APIs with RAG systems, significantly enhancing the capabilities of LLMs like GPT-4 and Llama-3 [38–40]. By leveraging LangChain, developers can seamlessly connect LLMs to external data sources, enabling dynamic retrieval of relevant information to augment the model's responses [41–43]. This integration is particularly beneficial in applications requiring up-to-date, specific, and accurate information, such as educational tools, customer support systems, and technical documentation.

LangChain creates a pipeline combining LLMs with embedding models and vector databases. When a user submits a query to an LLM, the model first converts this query into a numeric format known as an embedding. This embedding represents the query in a way that machines can process. The embedding is then compared to a vector database containing embeddings of a vast array of documents or information sources. This comparison helps identify the most relevant documents or data related to the query. Once the appropriate documents are identified, the information is fetched from these external sources. This process ensures that the most current and specific information is available to generate a response. The retrieved data is combined with the LLM's pre-existing knowledge to develop a comprehensive and accurate response to the user's query. The final output can also include citations or references to the retrieved information sources, enhancing trust and transparency.

2.4 RAG with LangChain

Figure 3 illustrates the workflow of the RAG system using LangChain. The process begins with a user posing a question, which is input into the system. The system performs an intelligent search involving vector similarity search within a knowledge graph to find relevant information. This information and the original question are then passed to LLM for processing. The LLM, represented by a parrot, processes the input to generate a coherent and accurate response. The final output, depicted by a lightbulb, represents the generated answer based on the provided documents and relevant information. This flowchart encapsulates the steps from initial query to final answer generation, showcasing the integration of intelligent search techniques and LLM capabilities to enhance the accuracy and relevance of the responses.

This study follows the steps to implement RAG with LangChain as described in Table 2.

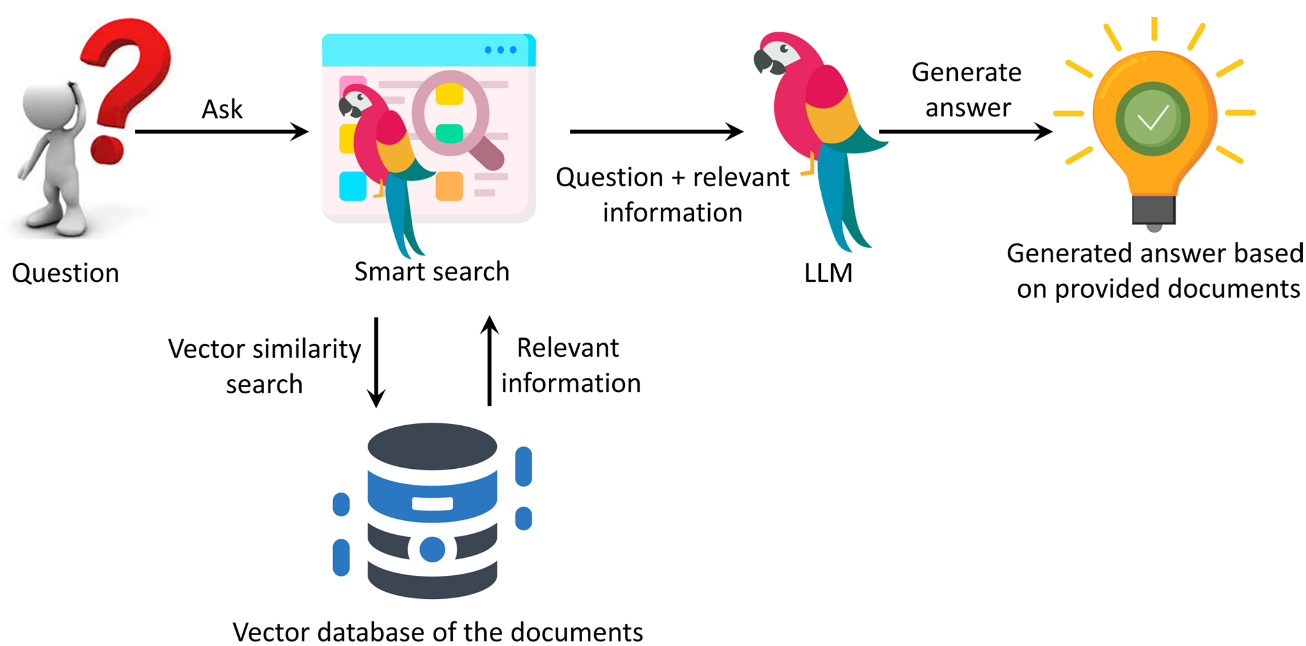


Fig. 3 Retrieval Augmented Generation (RAG) Flowchart demonstration

Table 2 Step-by-step implementation of Retrieval-Augmented Generation (RAG) using LangChain

Step	Description
Step 1: Environment Setup	A new virtual environment was created to manage the dependencies for the project. This ensures that all required libraries and packages are in a dedicated environment, avoiding conflicts with other projects
Step 2: Library Installation	The necessary libraries for the project were installed. These libraries include tools for loading and processing documents, creating embeddings, and building conversational AI models
Step 3: API Key Configuration	An OpenAI API key was required to access GPT-4. The API key was stored in a separate file named constants.py for security purposes. The key was then loaded into the environment, ensuring secure and easy access to the API
Step 4: Data Preparation	A directory named data was created to store the documents (PDF, DOCX, TXT) used by the chatbot. The script was designed to iterate through this directory and load the content of each file type using appropriate loaders
Step 5: Document Processing	The loaded documents were split into manageable chunks to facilitate efficient processing and retrieval. This was achieved using the CharacterTextSplitter class with chunk_size set at 200 and chunk_overlap at 10
Step 6: Vector Database Creation	A vector database was created from the document chunks using OpenAI embeddings, enabling efficient searching and retrieval of relevant document sections. FAISS was used for efficient similarity search
Step 7: Conversational Retrieval Chain Setup	A conversational retrieval chain was set up using the GPT-4 and Llama-3 models. The chain integrates a language model with a retrieval system and requires hyperparameters like temperature and search_kwargs
Step 8: Chatbot	A chatbot generates the response and formats it for the users

3 Results and discussions

3.1 Temperature settings in GPT API

The temperature parameter in LLMs like GPT-4 plays a pivotal role in defining the behaviour of generated text. Temperature controls the randomness in the model's responses by influencing the probability distribution over potential subsequent words [44, 45]. In practical applications, especially when using APIs, setting the appropriate temperature is crucial for optimising output quality [46].

At lower temperature settings, the model's outputs become more deterministic and focused [32]. This characteristic is particularly beneficial in technical and factual domains such as Geotechnical Engineering, where precision and reliability are paramount. By reducing randomness, the model adheres more strictly to the most probable sequences of words, thereby enhancing the accuracy of the information provided. Conversely, higher temperature settings introduce greater randomness, allowing the model to explore a broader range of responses. While this exploration can be advantageous for creative applications, it often leads to less accurate and relevant content due to the model generating text that diverges from core knowledge and retrieved information. This increased variability can introduce errors and reduce consistency. Moreover, at higher temperatures, there is an increased likelihood of hallucinations, where the model produces responses that are plausible in form but incorrect or irrelevant in substance.

3.2 Impact of temperature on accuracy

To assess the impact of temperature settings on the accuracy of LLM outputs, experiments were conducted with GPT-4 under Zero Shot and API configurations at three temperature settings: Temp = 0.1, Temp = 0.5, and Temp = 1. The top_p parameter was consistently set to its default value of 1, as altering both parameters simultaneously is not recommended [47]. The accuracy rates achieved for each configuration are illustrated in Fig. 4. The API with Temp = 0.1 setting improves the GPT-4's performance significantly from an accuracy rate of 0% under Zero Shot conditions to the highest accuracy at 95%, demonstrating that this low-temperature setting enhances performance by reducing output variability. The Temp = 0.5 setting shows a slightly lower accuracy at 82.5%, suggesting that moderate temperatures still maintain high accuracy but with some variability. The Temp = 1 setting results in a 60%

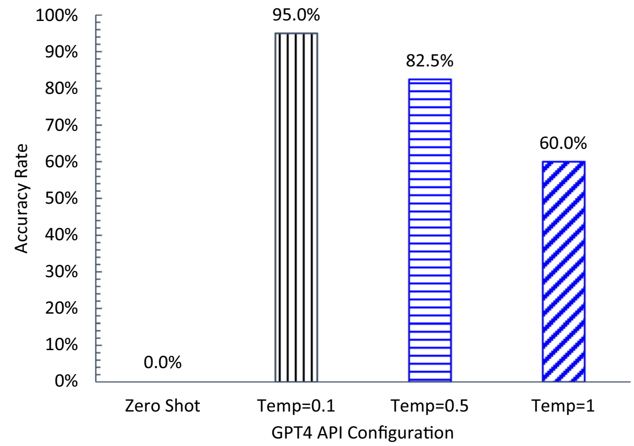


Fig. 4 Accuracy rate achieved by GPT-4 with the Zero Shot condition and GPT-4 API with different temperature settings

accuracy rate, the lowest among the temperature settings, indicating that higher temperatures introduce too much variability, reducing accuracy. Overall, GPT-4's accuracy improves with lower temperature settings. For high precision and factual correctness in applications like Geotechnical Engineering, a lower temperature setting is advisable to ensure reliable and effective outputs.

3.3 Comparison of GPT API and Llama API

Figure 5 illustrates the accuracy rates achieved by GPT-4 and Llama-3 under Zero Shot conditions and with API configurations, with temperature values set at 0.1. Under Zero Shot conditions, Llama-3 performs better than GPT-4, achieving an accuracy rate of 25% compared to GPT-4's 0%. Although this rate is still relatively low, it suggests that Llama-3 can manage some accuracy without prior input.

However, the improvements seen with API configurations show a slightly different trend. The Llama-3 API configuration shows an improvement, reaching an accuracy rate of 45%, up from 25% in its Zero Shot performance, marking a 20% increase. In contrast, the GPT-4 API configuration demonstrates a remarkable jump, soaring from 0% in Zero Shot to 95% correct with the API, indicating a 95% improvement. This substantial increase underscores the enhanced accuracy and reliability of the GPT-4 API when properly configured.

The figure highlights that GPT-4 and Llama-3 benefit significantly from API configurations, with notable improvements in their accuracy rates. However, GPT-4 shows a more substantial enhancement, reaching near-perfect accuracy with its API setup. This comparison suggests that API configurations are crucial for optimising the performance of these LLMs, particularly for tasks that require high precision and reliability. The outcome also implies that while both models improve with API configurations, GPT-4 has more potential for achieving higher accuracy in practical applications.

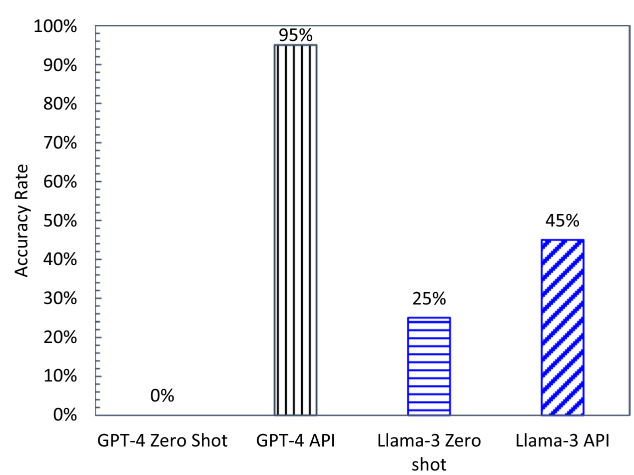


Fig. 5 Comparison of accuracy rate achieved by GPT-4 and Llama-3 with Zero Shot and API

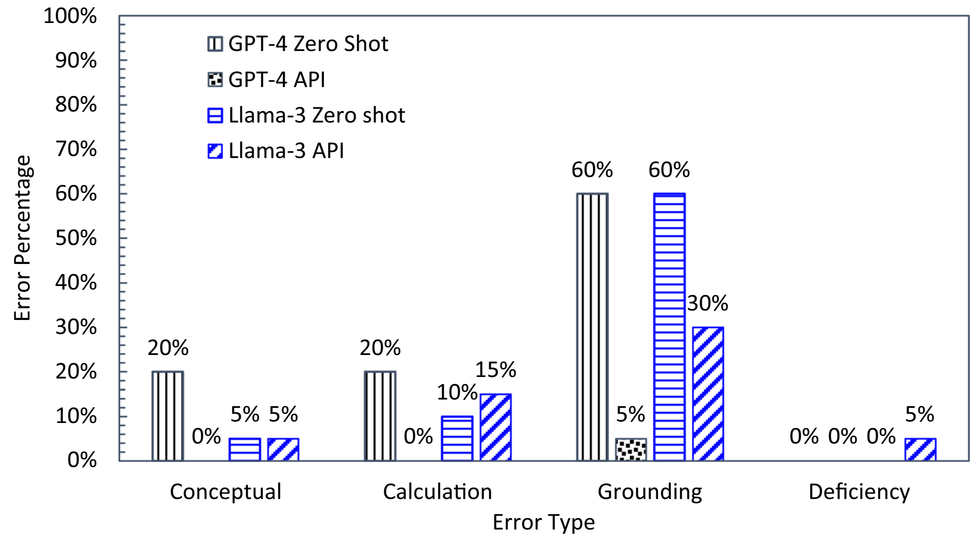


Fig. 6 Error percentage distribution by error type for GPT-4 and Llama-3 under Zero Shot and API configurations

Figure 6 displays the error percentage distribution across four error types for both GPT-4 and Llama-3 models under Zero Shot and API configurations. Regarding Conceptual Errors, GPT-4 exhibits a significant reduction when moving from Zero Shot to API configuration, dropping from 20 to 0%. This improvement underscores the model's enhanced ability to retrieve and apply necessary concepts with API support. Conversely, Llama-3 maintains a lower level of Conceptual Errors, at 5% in both Zero Shot and API configuration, indicating a relatively stable performance in this category. Calculation Errors also show a significant decrease for GPT-4, from 20% in Zero Shot to 0% in API configuration. This reduction highlights the model's improved accuracy in algebraic and arithmetic manipulations with API support. Llama-3, however, shows a slight increase in Calculation Errors from 10% in Zero Shot to 15% in API configuration, suggesting some challenges in maintaining computational accuracy under API settings. Grounding Errors are notably high in both models under Zero Shot conditions, with both GPT-4 and Llama-3 exhibiting a 60% error rate. With API support, these errors drastically decrease to 5% for GPT-4 and 30% for Llama-3. This significant reduction indicates that API configurations substantially enhance the models' ability to apply retrieved concepts correctly within equations or constraints. Deficiency Errors are absent in the Zero Shot configurations for both models but appear in the Llama-3 API configuration at a rate of 5%. This outcome suggests that while API support generally improves model performance, it may also introduce new challenges in interpreting images, graphs, and charts for Llama-3. Overall, the figure illustrates that both GPT-4 and Llama-3 benefit from API configurations, with GPT-4 showing a more pronounced improvement across all error types. The reduction in Grounding Errors is particularly notable, underscoring the critical role of API configurations in enhancing model performance. However, the emergence of Deficiency Errors in the Llama-3 API configuration highlights areas for further improvement, suggesting that while API configurations are beneficial, they also present new challenges that need to be addressed.

The comparative analysis of GPT-4 API and Llama-3 API highlights the importance of specialised formula integration in AI-driven educational tools for engineering disciplines. The RAG approach emerged as a promising method for enhancing the accuracy of AI responses, suggesting that the effectiveness of LLMs in educational settings can be significantly improved by leveraging external databases and structured knowledge.

Ultimately, the choice between GPT-4 API and Llama-3 API depends on the specific requirements and constraints of the application. GPT-4 is ideal for scenarios demanding high accuracy, nuanced understanding, and seamless cloud-based integration, offering reliable performance without substantial local hardware. On the other hand, Llama-3 provides a cost-effective, customisable solution suitable for targeted applications requiring specialised adjustments and efficient resource use. However, it necessitates an initial investment in hardware for local deployment.

To reinforce the reliability of the accuracy evaluation, we conducted statistical significance tests to assess differences between model performances. To quantify uncertainty in the accuracy estimates, we computed 95% confidence intervals (CIs) for each model using the Wilson Score Interval proposed by Wilson [48]. The results are summarised in Table 3.

To assess whether the observed differences in model accuracy are statistically significant, we conducted a chi-square test comparing GPT-4 (Temp = 0.1) and Llama-3 API.

Table 3 Accuracy rates of different LLM configurations in solving Geotechnical Engineering problems

Model & Setting	Accuracy (%)	95% CI Lower	95% CI Upper
GPT-4 (Temp = 0.1)	95.0	76.4	99.4
GPT-4 (Temp = 0.5)	82.5	61.4	93.8
GPT-4 (Temp = 1)	60.0	37.8	79.2
Llama-3 API	45.0	24.5	67.3
Llama-3 Zero-Shot	25.0	10.9	47.1

(1) Null Hypothesis (H_0): There is no significant difference in accuracy between GPT-4 (Temp = 0.1) and Llama-3 API.
(2) Alternative Hypothesis (H_1): There is a statistically significant difference in accuracy between these models.

The results of the chi-square test are summarised in Table 4.

Since the p-value (0.0019) is less than 0.05, we reject the null hypothesis and conclude that GPT-4 (Temp = 0.1) significantly outperforms Llama-3 API in accuracy. This analysis strengthens the validity of our findings and confirms that the performance differences observed in our study are statistically significant.

3.4 Tutor capabilities

In developing the AI tutor, we focused on the following key attributes to maximise its effectiveness in personalised education. The design objectives were to ensure adaptability to the learner's pace and preferences, enhance engagement, and maintain an environment conducive to diverse learning needs [34, 49].

(1) *Personalised Learning Experiences*: The AI tutor is engineered to tailor responses based on individual student profiles, which include their learning pace, preferences, and previous interactions. This customisation enhances the relevance and effectiveness of the educational content, making it more accessible to students from various skill sets.
(2) *Interactive Engagement*: By using clear and accessible language, the tutor maintains a high level of interaction. Engaging students in this manner helps to keep them motivated and improves retention of the material covered.
(3) *Support and Feedback*: Integral to the AI tutor's functionality is its ability to provide constructive feedback. By clearly explaining concepts and offering positive reinforcement, the tutor supports continuous learning and encourages academic exploration.
(4) *Resourcefulness and Flexibility*: The tutor is designed to suggest additional materials and alternative learning paths. This capability aids in the comprehension of complex topics and allows students to explore related areas independently, fostering a richer educational experience.
(5) *Safe and Ethical Learning Environment*: Adherence to privacy and ethical guidelines ensures that the learning environment remains secure, protecting student data and promoting a safe educational space.

These attributes are embedded within the prompt template that dictates the AI tutor's interactions with students, as demonstrated in Listing 1. This template serves as the operational foundation, ensuring that the AI's responses are accurate, informative, and specifically tailored to enhance the personalised learning experience. The effectiveness of this approach is supported by the following core educational theories:

(1) Constructivism

This theory, pioneered by Piaget [50] and Vygotsky and Luria [51], posits that learners construct knowledge actively rather than passively absorbing information. LLM-powered tutors align with this theory by facilitating interactive

Table 4 Chi-Square test results assessing the statistical significance of accuracy differences between AI models

Statistic	Value
Chi-square (χ^2)	9.64
p-value	0.0019

learning experiences where students can engage in problem-solving and inquiry-based learning. The adaptability of AI tutors allows personalised scaffolding, similar to Vygotsky's concept of the Zone of Proximal Development.

(2) Cognitive Load Theory

The Cognitive Load Theory was proposed by Sweller [52]. This theory emphasises the importance of managing intrinsic, extraneous, and germane cognitive loads to optimise learning. LLM-powered tutors can help mitigate extraneous load by simplifying complex explanations and tailoring responses to a learner's proficiency level. The RAG approach used in this study enhances information processing efficiency by integrating external resources, reducing cognitive overload.

(3) Self-Regulated Learning Theory (SRL)

Self-Regulated Learning Theory, which was proposed by Zimmerman [53] and Schunk and Zimmerman [54], highlights goal setting, self-monitoring, and metacognitive reflection. AI tutoring systems encourage self-regulated learning by providing immediate feedback, personalised hints, and adaptive questioning. The ability of LLM-powered tutors to track student progress and adjust their instructional approach aligns with SRL.

Table 5 presents sample interactions between the GPT-4 API and students, illustrating how the implemented prompt template from Listing 1 effectively facilitates these engagements.

Table 5 Sample interactions between GPT-4 API and user using the implemented prompt template from Listing 1

Prompted Question (i.e., Q12.10 from the book written by Das and Sobhan [35]):

In a consolidated-drained triaxial test on a clay, the specimen failed at a deviator stress of 124 kN/m^2. If the effective stress friction angle is known to be 31°, what was the effective confining pressure at failure?

GPT-4's Response:

Understanding the Problem

In a consolidated-drained triaxial test, the failure of a specimen occurs when a particular stress state, defined by the deviator stress and the confining pressure, meets the failure criterion dictated by the soil's friction angle. The deviator stress ($\Delta\sigma$) in a triaxial test is defined as the difference between the axial stress (σ_1) and the confining pressure (σ_3), or:

$\Delta\sigma = \sigma_1 - \sigma_3$

Given:

- Deviator stress ($\Delta\sigma$) = 124 kN/m^2; and—Effective stress friction angle (ϕ') = 31°

Calculating Effective Confining Pressure at Failure

To relate these variables to the Mohr–Coulomb failure criterion, we express the principal stresses at failure in terms of the angle of internal friction. From the theory, we know that:

$\sigma_1 = \Delta\sigma + \sigma_3$

The Mohr–Coulomb failure criterion for maximum and minimum principal stresses is given, in its simplest form (ignoring cohesion for purely frictional material), by:

$\sigma_1 = \sigma_3 \times tan^2(45° + \frac{\phi\prime}{2})$

Replacing (σ_1) in the equation gives us:

$\Delta\sigma + \sigma_3 = \sigma_3 \times tan^2(45° + \frac{\phi\prime}{2})$

From this, we can solve for (σ_3), the effective confining pressure at failure. Let's start rearranging this equation to isolate (σ_3)

Step-by-step Rearrangement

To isolate σ_3, we can modify the equation as follows:

$\Delta\sigma = \sigma_3 \times (tan^2\left(45° + \frac{\phi\prime}{2}\right) - 1)$

This equation will give us the value of σ_3 if we substitute the values of $\Delta\sigma$ and $\phi\prime$

Next Steps

- Calculate $tan^2\left(45° + \frac{\phi\prime}{2}\right)$. Could you go ahead and compute this value based on the given friction angle (ϕ') of 31°?

- Once you have that calculation, use it to find the effective confining pressure σ_3 using the rearranged formula

This calculation will incorporate the theory we've discussed, and I'm here to help guide you through each step if you need further assistance!

```
prompt_template = """
# As an AI tutor, your role is to provide personalised, engaging, and supportive
learning experiences. Keep the following qualities in mind:

# 1. Personalisation: Tailor your responses based on the student's needs, learn-
ing pace, and preferences.

# 2. Interactive and Engaging: Provide interactive and engaging explanations.

# 3. Accessibility: Ensure your responses are clear and accessible to students of
different backgrounds and abilities.

# 4. Feedback and Support: Offer constructive feedback and support, explaining
concepts clearly.

# 5. Motivational: Encourage and motivate the student with positive reinforce-
ment.

# 6. Resourceful: Provide additional resources and references when needed.

# 7. Flexible Learning Paths: Allow exploration of related topics and provide flex-
ible learning paths.

# 8. Contextual Understanding: Use context from previous interactions to pro-
vide relevant assistance.

# 9. Assessment and Evaluation: Ask questions to gauge understanding and pro-
vide detailed feedback.

# 10. Safety and Privacy: Ensure the conversation is safe and respects privacy.

{context}

# Question:

# {question}
"""
```

Listing 1 AI tutor prompt template

3.5 GPT API vs CustomGPT

Similar to GPT API, CustomGPT is an alternative feature offered by OpenAI, designed to enhance GPT's functionality by allowing the integration of external knowledge for more accurate responses. We also compared GPT API and CustomGPT as follows:

(1) Ease of Use: CustomGPT is more accessible to build and requires no code, while the GPT API offers more flexibility and customisation.
(2) Extended Knowledge and Tools: Both options provide extended knowledge and tools, such as data retrieval and code execution.
(3) User Interface: CustomGPT uses the familiar chat interface, while the GPT API allows for a customised user experience.

(4) Maintenance: Maintenance is more straightforward with CustomGPT, as OpenAI handles hosting and updates. The GPT API requires managing the application.
(5) Business Opportunities: The GPT API offers more business opportunities, such as providing paid tiers, while Custom GPTs may have revenue sharing in the future.
(6) Developer Experience: CustomGPT requires no coding and can be built entirely within the chat interface, while the GPT API requires creating an application and hosting the code externally.
(7) Classroom Impact: GPT API in classrooms is cost-effective for short-term use as users do not need a ChatGPT Plus account. CustomGPT can be shared with students if account credentials are provided, though this can be problematic due to query limits per hour in CustomGPT.

3.6 Ethical considerations

The deployment of AI-powered tutors in undergraduate Geotechnical Engineering and education broadly necessitates a thorough examination of the ethical implications associated with its use [55–57]. Data privacy, algorithmic bias, and the impact on traditional teaching roles are three primary areas of concern, as detailed below.

(1) Data Privacy: Using LLMs like GPT-4 and Llama-3 requires processing vast amounts of personal and academic data from students to tailor educational experiences effectively. Ensuring the confidentiality and security of this data is paramount. Transparency data handling and storage policies must be in place, adhering to regulations such as the General Data Protection Regulation (GDPR) in the European Union or Family Educational Rights and Privacy Act (FERPA) in the United States. Additionally, the consent of students and educators must be obtained, clearly outlining how their data will be used and the measures taken to protect it.
(2) Algorithmic Bias: AI models, including LLMs, are trained on large datasets that may contain historical biases. These biases can be inadvertently perpetuated and amplified by AI, leading to unfair or discriminatory educational experiences. Continuous monitoring and updating of AI models are crucial to identify and mitigate any embedded biases, ensuring that AI tutors provide equitable educational support to all students.
(3) Impact on Traditional Teaching Roles: Integrating AI tutors in education introduces concerns about the potential displacement of traditional teaching roles and the devaluation of human educators. It is crucial to frame AI tutors not as replacements but as supplements to human educators, enhancing the educational experience by providing additional support for personalised attention and immediate feedback. Educators play a crucial role in interpreting AI-generated content, providing context, and fostering critical thinking and emotional intelligence skills that AI cannot replicate.

3.7 LLM APIs limitations

The LLM APIs, while advanced, present their own set of limitations. GPT-4's impressive performance, particularly with the RAG approach, underscores the potential of integrating external knowledge sources. However, this reliance also highlights a critical limitation: the model's understanding and responses are as good as the data it can access, which may sometimes be outdated or biased. Furthermore, the RAG approach requires significant computational resources and sophisticated data management strategies, potentially limiting its accessibility for some educational institutions.

Llama-3's Zero Shot capabilities demonstrate the model's adaptability and the potential for wide application without extensive customisation. Yet, this strength is also a weakness, as the lack of domain-specific tuning can lead to inaccuracies or oversimplified explanations that might not meet the educational standards required for complex engineering concepts.

3.8 Study limitations and areas for AI tutor enhancement

The scope of this research is confined to undergraduate Geotechnical Engineering education, which may not fully represent the challenges and opportunities of using LLM APIs in other engineering disciplines or educational levels. Moreover, the comparative analysis primarily focused on two LLM APIs—GPT-4 and Llama-3—selected for their prominence and accessibility. This selection excludes other emerging LLMs that offer unique advantages or drawbacks in educational contexts. The study's methodology, emphasising formula integration and problem-solving, may also overlook other critical aspects of learning, such as fostering creativity, critical thinking, and collaboration among students.

One other limitation of this study is the exclusive use of text-based questions. In Geotechnical Engineering, many problems involve graphical data representations, such as Mohr's circles, soil classification charts, and stress–strain curves. Since no images or graphs were provided as input, the AI models were unable to demonstrate their ability to process visual information—a key aspect of real-world engineering problem-solving.

Future studies should explore multimodal AI models, such as GPT-4V (Vision) or Claude Opus, which can analyse both text and images. Evaluating LLM performance on graph-based geotechnical problems would provide deeper insights into their practical applicability in engineering education.

The limitations identified above point to several key areas for improvement in AI tutor development. Enhancing the quality and diversity of datasets used to train LLMs can address biases and ensure that AI tutors provide more accurate, inclusive, and contextually relevant support. Developing more sophisticated data integration and processing techniques can improve the models' ability to handle complex, formula-driven problems typical in engineering education.

Additionally, there is a pressing need for mechanisms that allow educators to easily update and customise AI tutors, ensuring that the content remains current with the latest scientific developments and pedagogical strategies. Finally, fostering interdisciplinary collaboration between AI researchers, educators, and students can lead to the development of more effective, user-friendly, and ethically responsible AI educational tools.

4 Conclusion and implications for future research

This study demonstrates that both GPT-4, as a commercial model, and Llama-3, as an open-source model, can effectively generate accurate answers and step-by-step explanations for typical Geotechnical Engineering problems. These capabilities suggest that LLMs may serve as a useful supplementary resource, akin to textbooks or solution manuals, for students seeking additional guidance outside the classroom. Llama-3 showed a reasonable understanding of the subject matter, albeit with slightly less precision in formula application than GPT-4. If Llama-3 is augmented with math capability, the accuracy might improve as it might reduce its deficiency in solving complex maths problems.

Our results also provide new insights into how LLMs perform on Geotechnical Engineering problems, particularly in relation to formula application, conceptual reasoning, and physics-based problem-solving. A notable finding of this research is that GPT-4 with Retrieval-Augmented Generation (RAG) achieved a 95% accuracy rate on formula-driven problems, demonstrating that LLMs can successfully retrieve and apply standard Geotechnical equations when properly prompted. However, accuracy declined when problems required higher-order conceptual reasoning, particularly when:

(1) Interpreting problem constraints from textual descriptions (e.g., determining which soil parameters to use when multiple is provided).
(2) Applying multi-step physical principles, such as stress–strain relationships and seepage analysis.
(3) Understanding implicit assumptions in soil behaviour, such as the role of effective stress versus total stress in consolidation problems.

The evaluation based on the set criteria—accuracy of content, formula integration, clarity of explanations, and adaptability—revealed that while no model perfectly addresses all educational needs in Geotechnical Engineering, each has unique strengths that contribute to their potential as AI tutors. This study demonstrates that GPT-4 with RAG can function as a supplementary learning tool for Geotechnical Engineering students, particularly in solving structured numerical problems with well-defined parameters. However, its limitations in contextual reasoning, multi-step dependency problems, and boundary condition interpretation suggest that AI tutors should be used as assistive tools rather than replacements for traditional instruction. Future work should explore how AI models can be optimised for conceptual understanding and decision-making in engineering applications.

Importantly, this study demonstrates that AI tutors can provide structured guidance and interactive problem-solving support, rather than merely offering solutions. By encouraging student engagement, prompting step-by-step calculations, and offering conceptual reinforcement, the AI tutor functions as a true learning aid rather than a static answer provider. The 24/7 availability of these AI tutors further ensures that students can access learning support at any time, fostering independent problem-solving and deeper conceptual understanding.

The potential of LLM APIs, such as GPT-4 and Llama-3, in transforming Geotechnical Engineering education is evident. As AI technology continues to evolve, so will its capabilities to support and enrich the learning journey of engineering students. By harnessing these advancements, educators can look forward to more personalised, engaging, and effective teaching tools tailored to the challenges of educating the next generation of engineers.

While this study highlights the promise of AI tutors in Geotechnical Engineering, several areas warrant further exploration:

(1) *Content Development*: The technique could also be used to develop exam questions, rubric development for assignments, role-playing scenarios such as Geotechnical professional evaluations, providing feedback on assignments, and brainstorming ideas.
(2) *Experimental Designs*: Future studies could employ longitudinal designs to assess the long-term effectiveness of AI tutors in Geotechnical Engineering education in the real-world environment. Tracking students' progress over semesters or years could provide deeper insights into how AI tutoring impacts knowledge retention, conceptual understanding, and practical application skills over time. Additionally, experimental designs incorporating control groups—where some students receive AI tutoring, and others do not—could offer more definitive conclusions about the efficacy of these tools.
(3) *Datasets*: Developing and validating specialised datasets for training LLMs in Geotechnical Engineering are crucial. Future research should focus on curating high-quality, diverse datasets that cover a broad spectrum of Geotechnical Engineering topics, including emerging areas of research and development. These datasets should also aim to be inclusive, representing a wide range of geographical regions, soil types, and engineering challenges to ensure the models' applicability across different contexts.
(4) *Formula Integration Techniques*: Exploring advanced formula integration techniques is another vital area for future research. This includes the development of novel algorithms that can better understand and process complex mathematical expressions, diagrams, and data tables common in Geotechnical Engineering. Investigating the integration of symbolic computation libraries with LLMs could enhance their ability to solve complex numerical problems accurately. Moreover, using augmented reality and virtual reality technologies to visualise complex geotechnical processes and simulations in tandem with AI tutors could revolutionise how students interact with and understand the subject matter.
(5) *Broader Educational Contexts*: Future studies should also consider the application of AI tutors in a broader range of educational settings, including vocational training programs, continuing education for professionals, and interdisciplinary engineering courses. Evaluating the effectiveness of AI tutors across different learning environments, such as online learning platforms, hybrid classrooms, and traditional lecture halls, could provide insights into how these tools can be best implemented to suit various educational needs.
(6) *Ethical and Social Implications*: Further research is needed to explore the ethical and social implications of deploying AI tutors in education. This includes studying the impact of AI on educational equity, student privacy, and the digital divide. Developing frameworks for ethical AI use in academic settings, addressing bias in AI algorithms, and ensuring that AI tutors are accessible to all students regardless of their socio-economic background are critical areas for investigation.

The findings of this study provide actionable insights for educators and institutions looking to integrate AI-driven tools into engineering education. By leveraging GPT-4 with RAG-based optimisation, educators can improve the accuracy and reliability of AI tutors, ensuring that students receive precise and context-aware explanations for Geotechnical Engineering problems. This approach can be particularly beneficial in flipped classroom models or blended learning environments, where AI tutors serve as on-demand learning assistants, allowing instructors to focus on higher-order conceptual discussions rather than repetitive problem-solving.

Additionally, the study highlights the potential for multimodal AI integration—combining text-based LLM responses with visual aids, diagrams, and interactive simulations—to enhance student engagement and comprehension. Educators can adopt these techniques by incorporating AI-generated explanations into digital learning platforms, integrating automated assessment tools, and using AI tutors for scaffolded learning experiences in STEM courses.

These findings highlight the growing role of AI in reshaping traditional engineering education, and future research should explore best practices for AI implementation, including pedagogical strategies for effectively incorporating AI tutors into real-world classrooms.

Acknowledgements During the preparation of this manuscript, the authors utilised GPT-4 to assist with editing. Subsequent to using this tool, the authors carefully reviewed and revised the content as necessary. Full responsibility for the final content of this publication rests with the authors.

Author contributions Conceptualisation, A.T., L.C., U.H. and J.K.; methodology, L.C., A.T., U.H. and J.K.; software, A.T., L.C. and U.H.; validation, A.T. and L.C.; formal analysis, L.C., A.T. and U.H.; investigation, L.C., A.T., U.H. and J.K.; resources, A.T.; data curation, L.C.; writing—original draft preparation, L.C. and A.T.; writing—review and editing, J.K.; visualisation, A.T., L.C. and U.H.; supervision, J.K.; project administration, J.K.; funding acquisition, J.K. All authors have read and agreed to the published version of the manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This research work was sponsored by the SPARC Hub (https://sparchub.org.au), which is affiliated with the Department of Civil Engineering at Monash University. This project is supported under the Australian Research Council (ARC) Industrial Transformation Research Hub (ITRH) Scheme (Project ID: IH180100010). We extend our sincere gratitude to Monash University and the SPARC Hub for their financial and in-kind support.

Data availability Data sets generated during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publications Not applicable.

Competing interests The authors declare no competing interests.

References

1. Harry A. Role of AI in education. Interdicip J Humman INJURITY. 2023. https://doi.org/10.58631/injurity.v2i3.52.
2. Kohnke L, Moorhouse BL, Zou D. ChatGPT for language teaching and learning. RELC J. 2023. https://doi.org/10.1177/00336882231162868.
3. Kumar K. Geotechnical parrot tales (GPT): harnessing large language models in geotechnical engineering. J Geotech Geoenviron Eng. 2024. https://doi.org/10.1061/JGGEFK.GTENG-11828.
4. Saputra I, Astuti M, Sayuti M, Kusumastuti D. Integration of artificial intelligence in education: opportunities, challenges, threats and obstacles: a literature review. Indones J Computer Sci. 2023. https://doi.org/10.33022/ijcs.v12i4.3266.
5. Song Y. Redefining STEM education in the post-ChatGPT era—case studies and perspectives. SSRN Electron J. 2024. https://doi.org/10.2139/ssrn.4733685.
6. Wang X, Li L, Tan SC, Yang L, Lei J. Preparing for AI-enhanced education: conceptualizing and empirically examining teachers' AI readiness. Computers Hum Behav. 2023. https://doi.org/10.1016/j.chb.2023.107798.
7. Pursnani V, Sermet Y, Kurt M, Demir I. Performance of ChatGPT on the US fundamentals of engineering exam: comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. Computers Educ Artif Intell. 2023. https://doi.org/10.1016/j.caeai.2023.100183.
8. VarastehNezhad A, Tavasoli R, Masumi M, Taghiyareh F. LLM performance assessment in computer science graduate entrance exams. In: 2024 11th International Symposium on Telecommunications (IST). IEEE; 2024. pp. 232–7.
9. Maitland A, Fowkes R, Maitland S. Can ChatGPT pass the MRCP (UK) written examinations? Analysis of performance and errors using a clinical decision-reasoning framework. BMJ Open. 2024. https://doi.org/10.1136/bmjopen-2023-080558.
10. Honig C, Rios S, Desu A. Generative AI in engineering education: understanding acceptance and use of new GPT teaching tools within a UTAUT framework. Australas J Eng Educ. 2025. https://doi.org/10.1080/22054952.2025.2467500.
11. Xu HR, Zhang N, Yin ZY, Guy Atangana Njock P. GeoLLM: a specialized large language model framework for intelligent geotechnical design. Computers Geotech. 2025. https://doi.org/10.1016/j.compgeo.2024.106849.
12. Wu S, Otake Y, Mizutani D, Liu C, Asano K, Sato N, Saito T, Baba H, Fukunaga Y, Higo Y, Kamura A, Kodama S, Metoki M, Nakamura T, Nakazato Y, Shioi A, Takenobu M, Tsukioka K, Yoshikawa R. Future-proofing geotechnics workflows: accelerating problem-solving with large language models. Georisk. 2024. https://doi.org/10.1080/17499518.2024.2381026.
13. Nguyen DV, Nguyen QN. Evaluating the symbol binding ability of large language models for multiple-choice questions in vietnamese general education. In: ACM International Conference Proceeding Series. Association for Computing Machinery. 2023; pp. 379–86.

14. Yang R, Yang B, Ouyang S, She T, Feng A, Jiang Y, Lecue F, Lu J, Li I. Leveraging large language models for concept graph recovery and question answering in NLP education. 2024.
15. Cain W. Prompting change: exploring prompt engineering in large language model AI and its potential to transform education. TechTrends. 2024;68(1):47–57. https://doi.org/10.1007/s11528-023-00896-0.
16. Lee U, Jung H, Jeon Y, Sohn Y, Hwang W, Moon J, Kim H. Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education. Educ Inf Technol. 2023. https://doi.org/10.1007/s10639-023-12249-8.
17. Chen L, Tophel A, Hettiyadura U, Kodikara J. An investigation into the utility of large language models in geotechnical education and problem solving. Geotechnics. 2024;4(2):470–98. https://doi.org/10.3390/geotechnics4020026.
18. Lu J, Yu L, Li X, Yang L, Zuo C. LLaMA-reviewer: advancing code review automation with large language models through parameter-efficient fine-tuning. In: Proceedings—International Symposium on Software Reliability Engineering, ISSRE. 2023.
19. Chen B, Yi F, Varró D. Prompting or fine-tuning? A comparative study of large language models for taxonomy construction. In: 2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C). IEEE; 2023. pp. 588–96.
20. Ge J, Sun S, Owens J, Galvez V, Gologorskaya O, Lai JC, Pletcher MJ, Lai K. Development of a liver disease–specific large language model chat interface using retrieval-augmented generation. Hepatology. 2024. https://doi.org/10.1097/hep.0000000000000834.
21. Jeong C. A study on the implementation of generative AI services using an enterprise data-based LLM application architecture. Adv Artif Intell Mach Learn. 2023. https://doi.org/10.54364/aaiml.2023.1191.
22. Thompson WE, Vidmar DM, De Freitas JK, Pfeifer JM, Fornwalt BK, Chen R, Altay G, Manghnani K, Nelsen AC, Morland K, Stumpe MC, Miotto R. Large language models with retrieval-augmented generation for zero-shot disease phenotyping. 2023.
23. Zhang Y, Zhu M, Gong Y, Ding R. Optimizing science question ranking through model and retrieval-augmented generation. Int J Computer Sci Inform Technol. 2023. https://doi.org/10.62051/ijcsit.v1n1.17.
24. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. 2020.
25. Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, Gao J. Large language models: a survey. 2024.
26. Raiaan MAK, Mukta MSH, Fatema K, Fahad NM, Sakib S, Mim MMJ, Ahmad J, Ali ME, Azam S. A review on large language models: architectures, applications, taxonomies, open issues and challenges. IEEE Access. 2024;12:26839–74. https://doi.org/10.1109/ACCESS.2024.3365742.
27. Xu X, Li M, Tao C, Shen T, Cheng R, Li J, Xu C, Tao D, Zhou T. A survey on knowledge distillation of large language models. 2024.
28. Chai SH, Chen I, Huang J, Yacoub T. Large language model for geotechnical engineering applications using retrieval augmented generation. In: Geotechnical frontiers. Reston: American Society of Civil Engineers; 2025. pp. 1–10.
29. Krathwohl DR. A revision of bloom's taxonomy: an overview. Theory Into Practice. 2002;41(4):212–8. https://doi.org/10.1207/s15430421tip4104_2.
30. Ranaldi L, Freitas A. Aligning large and small language models via chain-of-thought reasoning. In: 18th Conference of the European Chapter of the Association for Computational Linguistics. Long Papers. 2024. pp. 1812–27.
31. Hadi MU, Qureshi R, Shah A, Irfan M, Zafar A, Shaikh MB, Akhtar N, Wu J, Mirjalili S. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. Authorea Preprints. Authorea. 2023.
32. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep. 2023. https://doi.org/10.1038/s41598-023-46995-z.
33. Chiasson RM, Goodboy AK, Vendemia MA, Beer N, Meisz GC, Cooper L, Arnold A, Lincoski A, George W, Zuckerman C, Schrout J. Does the human professor or artificial intelligence (AI) offer better explanations to students? Evidence from three within-subject experiments. Commun Educ. 2024;73(4):343–70. https://doi.org/10.1080/03634523.2024.2398105.
34. Afzal S, Dempsey B, D'Helon C, Mukhi N, Pribic M, Sickler A, Strong P, Vanchiswar M, Wilde L. The personality of ai systems in education: experiences with the watson tutor, a one-on-one virtual tutoring system. Child Educ. 2019. https://doi.org/10.1080/00094056.2019.1565809.
35. Das B, Sobhan K. Principles of Geotechnical engineering, Eight Edition. In: Cengage Learning. 2014.
36. OpenAI. GPT-4 Technical Report. 2023;4: 1–100.
37. OpenAI. Models—OpenAI API. 2023.
38. Nascimento E, García G, Victorio W, Lemos M, Izquierdo Y, Garcia R, Paes Leme LAP, Casanova MA. A family of natural language interfaces for databases based on ChatGPT and LangChain. In: CEUR Workshop Proceedings. 2023.
39. Neira-Maldonado P, Quisi-Peralta D, Salgado-Guerrero J, Murillo-Valarezo J, Cárdenas-Arichábala T, Galan-Mena J, Pulla-Sanchez D. Intelligent educational agent for education support using long language models through Langchain. In: Lecture Notes in Networks and Systems. 2024.
40. Soygazi F, Oguz D. An analysis of large language models and LangChain in mathematics education. In: ACM International Conference Proceeding Series. 2023.
41. Auffarth B. Generative AI with LangChain: build large language model (LLM) apps with Python, ChatGPT, and other LLMs. Birmingham: Packt Publishing; 2023.
42. Topsakal O, Akinci TC. Creating large language model applications utilizing LangChain: a primer on developing LLM apps fast. Int Conf Appl Eng Nat Sci. 2023. https://doi.org/10.59287/icaens.1127.
43. Workman AD, Rathi VK, Lerner DK, Palmer JN, Adappa ND, Cohen NA. Utility of a LangChain and OpenAI GPT-powered chatbot based on the international consensus statement on allergy and rhinology: rhinosinusitis. Int Forum Allergy Rhinol. 2023. https://doi.org/10.1002/alr.23310.
44. Antaki F, Milad D, Chia MA, Giguère CÉ, Touma S, El-Khoury J, Keane PA, Duval R. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. Br J Ophthalmol. 2023. https://doi.org/10.1136/bjo-2023-324438.

45. Civettini I, Zappaterra A, Ramazzotti D, Granelli BM, Rindone G, Aroldi A, Bonfanti S, Colombo F, Fedele M, Grillo G, Parma M, Perfetti P, Terruzzi E, Gambacorti-Passerini C, Cavalca F. Evaluating the performance of large language models in hematopoietic stem cell transplantation decision making. Blood. 2023. https://doi.org/10.1182/blood-2023-185854.
46. Wang C, Liu SX, Awadallah AH. Cost-effective hyperparameter optimization for large language model generation inference. In: Proceedings of Machine Learning Research. 2023.
47. OpenAI. OpenAI Platform. 2024. https://platform.openai.com/docs/api-reference/chat/create. Accessed 6 Oct 2024.
48. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927. https://doi.org/10.1080/01621459.1927.10502953.
49. Kim WH, Kim JH. Individualized AI tutor based on developmental learning networks. IEEE Access. 2020. https://doi.org/10.1109/ACCESS.2020.2972167.
50. Piaget J. To understand is to invent: the future of education. London: Penguin Books; 1976.
51. Vygotsky L, Luria A. Tool and symbol in child development. Mind in Society. 1978.
52. Sweller J. Cognitive load during problem solving: effects on learning. Cognit Sci. 1988. https://doi.org/10.1016/0364-0213(88)90023-7.
53. Zimmerman BJ. Becoming a self-regulated learner: an overview. Theory Pract. 2002. https://doi.org/10.1207/s15430421tip4102_2.
54. Schunk DH, Zimmerman BJ. Motivation and self-regulated learning: theory, research, and applications. In: Schunk DH, Zimmerman BJ, editors. Motivation and self-regulated learning: theory, research, and applications. Milton Park: Routledge; 2012.
55. Guleria A, Krishan K, Sharma V, Kanchan T. ChatGPT: ethical concerns and challenges in academics and research. J Infect Dev Ctries. 2023. https://doi.org/10.3855/jidc.18738.
56. Liyanage UP, Ranaweera ND. Ethical considerations and potential risks in the deployment of large language models in diverse societal contexts. J Comput Soc Dyn. 2023;8(11):15–25.
57. Vidhya NG, Devi D, Nithya A, Manju T. Prognosis of exploration on Chat GPT with artificial intelligence ethics. Braz J Sci. 2023. https://doi.org/10.14295/bjs.v2i9.372.