ŌTAUTAHI CHRISTCHURCH AOTEAROA NEW ZEALAND
35TH AUSTRALASIAN ASSOCIATION FOR ENGINEERING EDUCATION ANNUAL CONFERENCE
8-11 DEC 2024
THE ENGINEER AND THE WORLD

Use of Large Language Models in Engineering Education: A Case Study on Infrastructure Design Report Introductions

Linxin Hua[a*]; Nan Zheng[a], Ye Lu[a], Lirui Guo[a], Jia Xu[b].
Department of Civil Engineering, Monash University, Melbourne, VIC, Australia[a]
Department of Information Systems and Business Analytics, Deakin University, Melbourne, VIC, Australia[b]
Corresponding Author Email: linxin.hua@monash.edu

ABSTRACT

CONTEXT

This paper presents a case study focusing on the application of Large Language Models (LLMs) in engineering learning: composing the introduction section of a design report. Utilising three state-of-the-art (SOTA) LLMs, this study explores their capabilities through simple prompt and zero-shot Chain of Thought (CoT), examines common errors in the generations, and proposes an ad-hoc generation strategy to enhance output quality based on these observations. This study aims to provide a case to demonstrate the practical use of LLMs to students, fostering a deeper understanding and more skilled application in their coursework and professional tasks.

PURPOSE

This research aims to develop a case study that can be applied in educational settings to effectively guide students on how to utilise LLMs, addressing the lack of practical resources in current curricula.

APPROACH

This study evaluated the performance of three SOTA LLMs to generate the introduction section of an infrastructure design report, employing simple prompt and the zero-shot COT prompting method. By analysing errors and limitations in the model outputs, this study proposes a tailored generation strategy to enhance the use of LLMs in introduction generation. This strategy aims to improve the accuracy and relevance of generated content in the task of generating introductions.

OUTCOMES

The study analysed outputs from three SOTA LLMs in generating introductions for engineering reports, identifying two key limitations: insufficient extraction and inadequate understanding of information under complex scenarios. To address these, it introduced a generation strategy where LLMs first process supporting materials to simplify and extract the key information, then generate the introduction based on this processed information, resulting in improved accuracy and stability.

SUMMARY

With a lack of instructions on effectively using LLMs in engineering education, this study introduces a case study for a typical engineering task. It discusses the limitations of SOTA models and popular prompting methods, proposing an effective generation strategy that shows improved performance. This case has the potential to serve as a practical guide for students, showcasing how to effectively employ LLMs to enhance problem-solving in engineering contexts.

KEYWORDS

Large Language Models in Education, Engineering Education, Prompt Engineering

Introduction

Large Language Models (LLMs) have already made a notable impact on tertiary education, with numerous initiatives aiming at incorporating them into the teaching process (Fütterer et al., 2023; Kasneci et al., 2023; Krause et al., 2024). For example, researchers have attempted to leverage the text-generation capabilities of LLMs to automate assignment grading (Moore et al., 2022) and provide personalised feedback (Liang et al., 2024). LLMs have also been used to develop learning assistant systems that aid students by providing timely responses to their questions (Caccavale et al., 2024; Tack & Piech, 2022). Moreover, researchers are exploring the capabilities of LLMs in understanding teaching materials and their performance in related assessments (L. Chen et al., 2024; Katz et al., 2023).

On the other hand, students are increasingly utilising LLMs in their learning processes (Zhang et al., 2024). However, current engineering education practices seldom provide guidance on the effective usage of LLMs in diverse tasks. In this case, student usage of LLMs is often more intuitive, which leads to their reliance on basic functionalities such as searching, translation, and text polishing (Zhang et al., 2024). Moreover, there is a lack of adequate awareness of the limitations of LLMs, which might cause inappropriate usage (Niedbał et al., 2023). Therefore, to maximise the utility of LLMs while mitigating their limitations, it is crucial to proactively equip students with robust methodologies and case studies that illustrate how to utilise LLMs effectively (Essel et al., 2024; Tsai et al., 2023).

However, in current engineering practices, the limited applications of LLMs result in a corresponding lack of practical cases in engineering education (Rane, 2023). Although some research has explored the capabilities of LLMs in understanding and solving geotechnical problems (B. Chen et al., 2024), this focus remains primarily on problem-solving rather than providing students with strategies for adopting LLMs in their studies. This problem-centric approach lacks the necessary context-based guidance. In a roundtable discussion which gathered learning and teaching leaders and students from 28 higher education institutions in Australia and New Zealand, Liu et al. (2023) pointed out the importance of equity and access, as well as normalising generative AI in higher education. They identified the lack of AI literacy as a significant barrier, highlighting the necessity to offer students training to democratise knowledge around generative AI. This lack of practical examples and strategies significantly impacts the ability to instruct students in the effective use of AI in engineering contexts.

This paper presents a detailed case study in response to the demands of engineering education. This case study focuses on a practical learning task, introduction generation for an infrastructure design report using various LLMs, including ChatGPT-4 (OpenAI et al., 2024), Claude 3-Opus (Anthropic, 2024), and Gemini 1.5 pro (Gemini Team Google et al., 2024). It evaluates the performance of these models and the widely-used prompting method, zero-shot Chain-of-Thought (CoT) (Chu et al., 2024; Wei et al., 2023) are discussed. Following an analysis of these generations, this study proposes a generation strategy to effectively improve the performance of LLMs in such scenarios. This strategy demonstrates its effectiveness in engineering report writing, highlighting the importance of analysing and optimising prompts.

Generation Task - Introduction section

Task background

This study uses the generation of the introduction section for an infrastructure design report to illustrate the possible limitations of using LLMs. This design report is part of a project-based assignment in the course CIV3283, a third year engineering core unit at Monash University. It is simplified from a project proposed by the local government to construct two roads connecting multiple locations, aiming to improve the accessibility of a logistics centre. The surveys of geological and traffic conditions have been completed. The teaching team has broken down the project into three design phases as shown in Figure 1 and mapped the tasks for all phases. All the available information, the project background, design objectives, site conditions, and expected deliverables, has been integrated into a 2000-word project brief, which is the only available input for LLMs.

The task is to generate the introduction section for the preliminary stage report, the first of the three design reports of this project. It is a typical descriptive task in engineering education, informing the readers of the key information about the project and the report. Therefore, an introduction section that meets the expectations should be a concise summary of the project and fully present the scope of the project and the report. In order to reduce irrelevant information and avoid exposing information about the project itself, only structures of the generated introductions are presented in this study, all sensible information has been anonymised.

Figure 1: Three phases of the infrastructure design project

Prompt engineering

When users interact with LLMs, a common approach is to provide a direct and simple prompt, such as “Please help me write a leave request email for the following reasons...”. This process gives LLMs straightforward instructions without examples to guide the generation. This prompting method is referred to as 'simple prompt' in this paper. In daily usage, this method may be sufficient to generate satisfactory results. However, for more complex scenarios in which the understanding and reasoning capabilities of LLMs are tested, simple prompts often perform poorly (B. Chen et al., 2024; Sahoo et al., 2024).

Research indicates that prompting LLMs to engage in reasoning or related activities improves the accuracy of their generated results (Sahoo et al., 2024). One commonly used method is 'Chain of Thought' (CoT) (Wei et al., 2023). This method enhances the accuracy of the outputs by incorporating reasoning steps into the prompts or instructing the LLMs to add reasoning processes. This guides the LLMs to mimic human thought processes in systematically analysing and reasoning through problems. Under zero-shot scenarios that have no instructive examples, this prompting method is achieved by appending 'let’s think step by step' to the original prompts. Due to its ease of use and the improvement it brings to any prompt, zero-shot CoT has become a popular prompting method.

A common practice is to provide human feedback to modify the generation, which relies on the subjective judgment and expertise of the users. Given the difference in various projects, it is impossible to reuse feedback and modification. Moreover, when feedback is provided by human users, it is hard to consider that the iterated generation reflects the capability of LLMs. Therefore, the generations in this study are produced without iteration and human intervention.

To simulate the everyday usage scenarios of students, this study will utilise the prompting methods that are likely the most common approaches: simple prompt and zero-shot Chain of Thought (CoT) to generate the introduction section.

Large Language Models for generation

Among the currently available LLMs, ChatGPT-4, Claude 3-Opus, and Gemini 1.5 pro are three of the most widely used and advanced models, outperforming most other LLMs across multiple datasets (Anthropic, 2024; Gemini Team Google et al., 2024; OpenAI et al., 2024). This study investigates the capabilities of these three LLMs in generating the introduction section of infrastructure design reports under specific project backgrounds. Due to the randomness of the models, the generated introduction sections vary among different runs. However, the identified errors are representative and continuously appear across multiple generations. All generations were collected between 31st March 2024 and 2nd April 2024.

Generation of the introduction section

Simple prompt

The ‘User’ section of Figure 2 presents the ‘simple prompt’ given to the LLMs. The prompt includes a request and the complete project brief. The remainder of Figure 2 shows the structures of the introductions generated by the selected LLMs. All three models partially achieved the objectives of an introduction section by generating content related to the project background, motivation, goal, and design principles, introducing the project to the readers.

However, the generated introduction sections of Stage 1 were insufficient across all three models. Gemini 1.5 Pro generated an overview of the Stage 1 report but overlooked part of the tasks in the Stage 1 design. ChatGPT-4 ignored the fact that this generation should summarise the Stage 1 design in the report. It is notable that Claude-3-Opus provided a list of Stage 1 tasks but only simply repeated the task names without further refinement or summarising of the information.

Additionally, some of the information generated from ChatGPT-4 and Claude-3-Opus was redundant for an introduction section. ChatGPT-4 repeated the detailed site information from the project brief, and Claude-3-Opus listed the design considerations for the alignment design. Both generations involved details that are incongruous with other parts of the introduction. Generally speaking, these specifics are typically covered in the corresponding sections rather than at the beginning of the report.

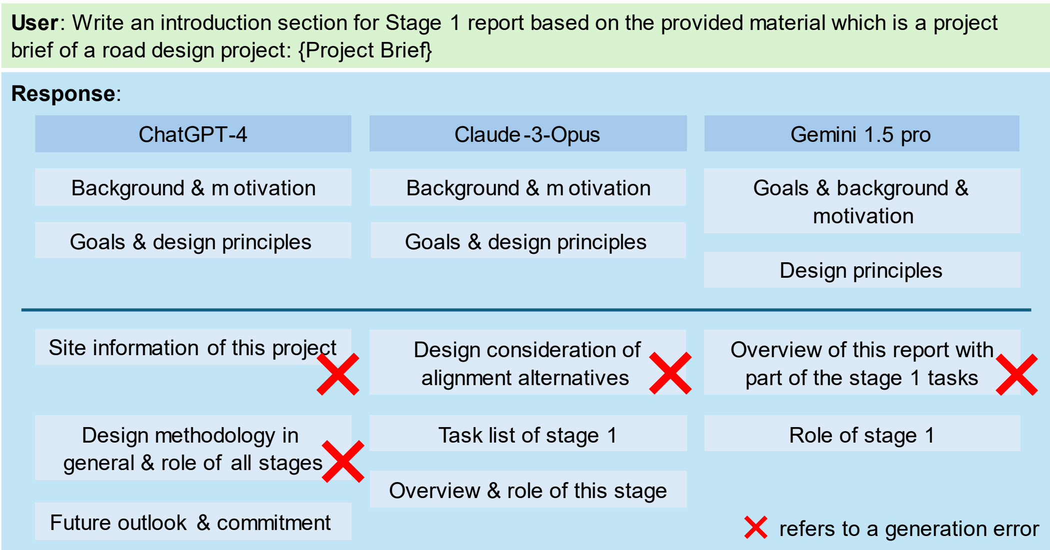

Figure 2: Simple prompt and structures of generated introductions

Zero-shot Chain of Thought

Zero-shot CoT adds “Let’s think step-by-step” after the original prompt, aiming to guide the LLMs to solve the problems progressively. Figure 3 shows the prompt used and the generated responses. Unexpectedly, apart from Claude-3-Opus, the other two models did not provide a step-by-step task analysis. Except that, similar to the simple prompt approach, all three models introduced the project well but overlooked some of the tasks in Stage 1 design. Compared to using a simple prompt, employing zero-shot CoT in ChatGPT-4 showed some improvement by providing specific information related to Stage 1. However, like the other two models, it only covered part of the design tasks of Stage 1. Additionally, Claude-3-Opus listed the geological investigations and project division which were completed in the project brief as Stage 1 design tasks, and Gemini 1.5 Pro mistakenly described the structure of the project brief as the structure of the Stage 1 report. Both resulted in factually incorrect statements.

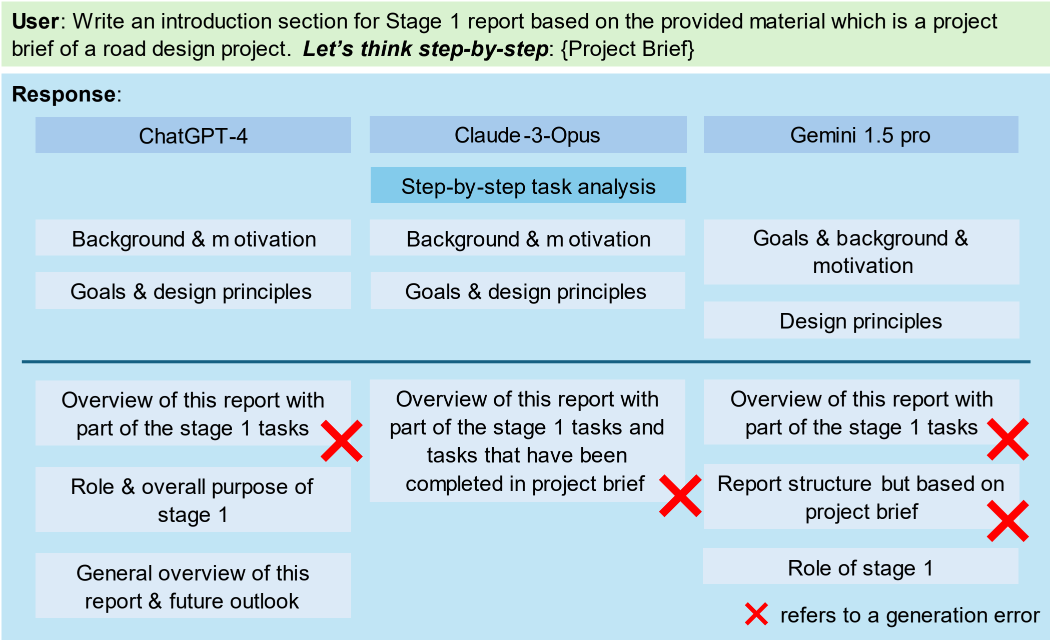

Figure 3: Zero-shot CoT prompt and structures of generated introductions

Problems in generation

It is noted that some generated results included the information from the project brief that was unrelated to the introduction, failing to recognise that the task was to generate the introduction section for the Stage 1 report. This indicates that in these generations, the LLMs only reorganised the content provided in the project brief and presented it in the generations without fully understanding the relationship between the information in the project brief and the requirements of the generation task. Additionally, it is worth noting that only Claude-3 Opus responded as expected under the zero-shot CoT prompting method, producing a step-by-step analysis.

The behaviour and performance of the LLMs suggest that when dealing with complex generation materials (e.g., a project brief), LLMs are likely to ignore or partially ignore some of the prompt instructions. It is challenging for the models to comprehensively extract the information and judge the importance of different information for a specific generation task. This shows the limited ability of LLMs to extract the necessary information from complex material.

Overall, the generation errors of LLMs when dealing with complex information and procedures can be summarised as follows:

1. **Insufficient extraction of information**

With complex supporting materials, some requests and critical information might be overlooked, leading to generations that do not fully address the prompts.

2. **Insufficient understanding of information**

For complex scenarios, the reasoning ability might be insufficient to understand the relationship between various parts of the supporting materials and clearly illustrate the correct relationship in one generation. Consequently, some information might be used inappropriately.

Ad-hoc generation strategy and results

Based on the behaviour of LLMs in generating introduction sections, it can be hypothesised that under the given materials and prompts, LLMs cannot adequately extract key information or understand the relationships between them. Therefore, this study proposes a generation strategy that divides the whole introduction into two generation tasks: (1) requiring LLMs to process the information provided by users and (2) generating the introduction section based on the processed information.

The prompt used for the first generation task is: *"Extract the key information from the project brief of a road design project, and present it in a multi-level list format."* In response to this prompt, LLMs are expected to organise and extract the key information, omitting details that are excessively redundant for the introduction section. Since information extraction itself is more fundamental than introduction generation, it is likely that the LLMs can meet expectations for this task. The prompt also requires LLMs to present the extracted information in a multi-level list format, which effectively simplifies the information for LLMs to identify the relationships among pieces of information. The purpose of this prompt is to instruct LLMs to simplify the project brief and highlight the importance and hierarchy of different information. This approach helps LLMs demonstrate the correct organisation and arrangement of content in the second generation task, which involves generating the introduction.

The prompts and key generation results of the three selected models are shown in Figure 4. All three models generated a multi-level list as expected. The lists produced by ChatGPT-4 and Claude-3-Opus were very similar, whereas the one generated by Gemini 1.5 Pro provided a more concise overview of the entire project, omitting detailed descriptions of the project. In the second generation task, all three models generated the introduction with similar structures. Moreover, they all avoided overemphasising detailed information in the project background section and emphasised the tasks covered in the Stage 1 report.

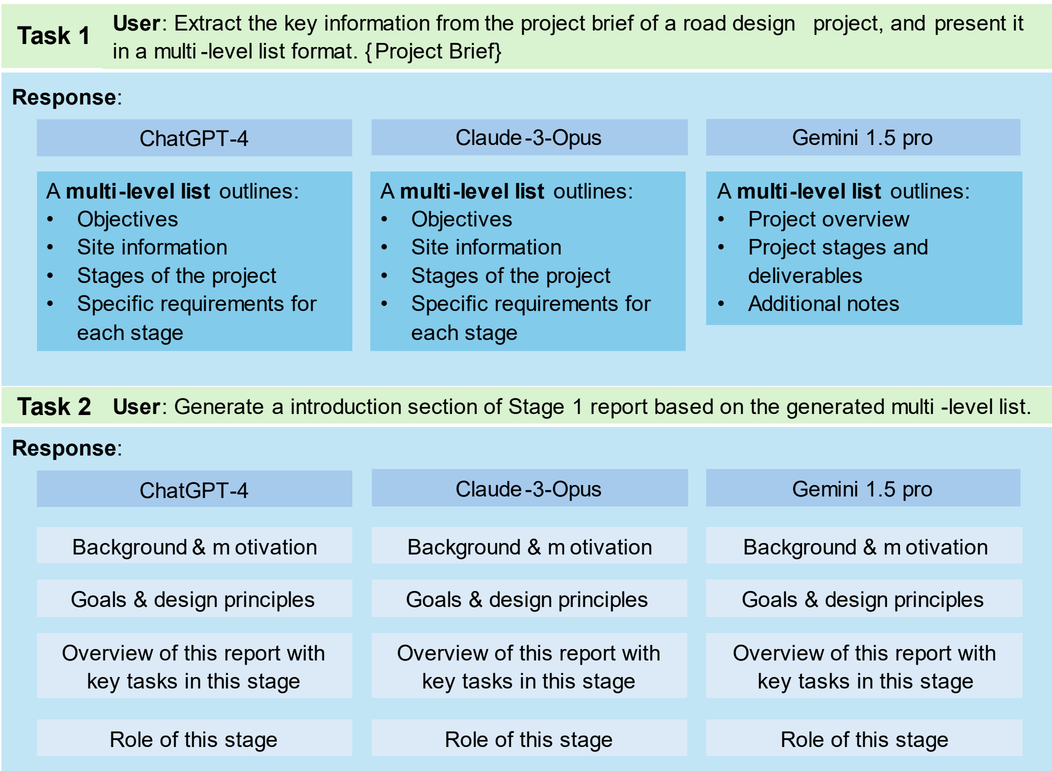

Figure 4: Prompts using the ad-hoc generation strategy and structures of generated introductions

It is noted that after processing the project brief, the Task 2 outputs that based on the restructured information from Task 1 were remarkably stable. All three models generated introductions with very similar structures, consistently aligning with the requirements of the introduction section and effectively eliminating the redundant details and irrelevant information commonly seen with simple prompt and zero-shot CoT approaches. Therefore, it can be concluded that the proposed generation strategy enhances the ability of LLMs to extract and utilise information for generating the introduction section in infrastructure design reports, thereby increasing the accuracy of the generated results. This approach can also help students better understand the tasks and validate the generated structures by the LLMs, instead of simply copying the generated content.

Conclusion

In current engineering education, there is a lack of methods and resources to guide students in using LLMs effectively. This study proposed an ad-hoc generation strategy that divides the generation tasks into two phases. By employing a case study of generating an introduction section of a design report using three state-of-the-art LLMs, this study demonstrates the outperformed performance of the ad-hoc generating strategy compared to simple prompt and zero-shot CoT approaches. This case study can be used in engineering education practice to guide students in using LLMs more effectively, demonstrating how to solve real tasks they may encounter in using LLMs during their learning process. The major contributions are as follows:

1. This study employs three state-of-the-art LLMs, ChatGPT-4, Claude 3-Opus, and Gemini 1.5 Pro, which have demonstrated superior performance across multiple databases. These models generated the introduction sections using both simple prompt and zero-shot CoT prompting methods. Based on the generated results, the study analyses and highlights their shortcomings in extracting and utilising information under engineering design report writing scenarios.
2. In response to these shortcomings, this study proposes a strategy of reorganising the project brief and then generating the introduction based on this restructured information. This generation strategy showed superior performance compared to the other two prompting methods on all three models. Without the input of human feedback, the randomness of the generations decreased, and the results more closely aligned with the requirements of the introduction section.

However, this study is subject to certain limitations, which motivate future work in the following areas:

1. While this study explores a practical application of LLMs in educational settings, further comprehensive investigation is needed to understand the influence of this case study under practical education settings. This case study will be presented to students as part of the teaching. The future research will further explore the integration of these case studies and teaching process to guarantee effective delivery of learning materials.
2. This study focuses on a specific task. However, engineering students may encounter various scenarios and types of tasks in their learning. The applicability of the analysis methods and generation strategies demonstrated in this study to other tasks requires further exploration.

References

Anthropic. (2024, May). Responsible Scaling Policy Evaluations Report – Claude 3 Opus. https://cdn.sanity.io/files/4zrzovbb/website/210523b8e11b09c704c5e185fd362fe9e648d457.pdf

Caccavale, F., Gargalo, C. L., Gernaey, K. V., & Krühne, U. (2024). Towards Education 4.0: The role of Large Language Models as virtual tutors in chemical engineering. Education for Chemical Engineers, 49, 1–11. https://doi.org/10.1016/j.ece.2024.07.002

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2024). Unleashing the potential of prompt engineering in Large Language Models: A comprehensive review (arXiv:2310.14735). arXiv. http://arxiv.org/abs/2310.14735

Chen, L., Tophel, A., Hettiyadura, U., & Kodikara, J. (2024). An Investigation into the Utility of Large Language Models in Geotechnical Education and Problem Solving. Geotechnics, 4(2), 470–498. https://doi.org/10.3390/geotechnics4020026

Chu, Z., Chen, J., Chen, Q., Yu, W., He, T., Wang, H., Peng, W., Liu, M., Qin, B., & Liu, T. (2024). Navigate through Enigmatic Labyrinth A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future (arXiv:2309.15402). arXiv. http://arxiv.org/abs/2309.15402

Essel, H. B., Vlachopoulos, D., Essuman, A. B., & Amankwa, J. O. (2024). ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). Computers and Education: Artificial Intelligence, 6, 100198. https://doi.org/10.1016/j.caeai.2023.100198

Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M., & Gerjets, P. (2023). ChatGPT in education: Global reactions to AI innovations. Scientific Reports, 13(1), 15310. https://doi.org/10.1038/s41598-023-42227-6

Gemini Team Google, Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., & Pan, Z. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (arXiv:2403.05530). arXiv. https://doi.org/10.48550/arXiv.2403.05530

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2023). GPT-4 Passes the Bar Exam. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4389233

Krause, S., Panchal, B. H., & Ubhe, N. (2024). The Evolution of Learning: Assessing the Transformative Impact of Generative AI on Higher Education (arXiv:2404.10551). arXiv. http://arxiv.org/abs/2404.10551

Liang, Z., Sha, L., Tsai, Y.-S., Gašević, D., & Chen, G. (2024). Towards the Automated Generation of Readily Applicable Personalised Feedback in Education. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), Artificial Intelligence in Education (Vol. 14830, pp. 75–88). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64299-9_6

Liu, D., Fawns, T., Cowling, M., & Bridgeman, A. (2023). Working paper: Responding to Generative AI in Australian Higher Education. https://doi.org/10.35542/osf.io/9wa8p

Moore, S., Nguyen, H. A., Bier, N., Domadia, T., & Stamper, J. (2022). Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3. In I. Hilliger, P. J. Muñoz-Merino, T. De Laet, A. Ortega-Arranz, & T. Farrell (Eds.), Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption (Vol. 13450, pp. 243–257). Springer International Publishing. https://doi.org/10.1007/978-3-031-16290-9_18

Niedbał, R., Sokołowski, A., & Wrzalik, A. (2023). Students' Use of the Artificial Intelligence Language Model in their Learning Process. Procedia Computer Science, 225, 3059–3066. https://doi.org/10.1016/j.procs.2023.10.299

OpenAI, 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., … Zoph, B. (2024). GPT-4 Technical Report (arXiv:2303.08774). arXiv. http://arxiv.org/abs/2303.08774

Rane, N. (2023). ChatGPT and Similar Generative Artificial Intelligence (AI) for Building and Construction Industry: Contribution, Opportunities and Challenges of Large Language Models for Industry 4.0, Industry 5.0, and Society 5.0. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4603221

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications (arXiv:2402.07927). arXiv. http://arxiv.org/abs/2402.07927

Tack, A., & Piech, C. (2022). The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues (arXiv:2205.07540). arXiv. http://arxiv.org/abs/2205.07540

Tsai, M.-L., Ong, C. W., & Chen, C.-L. (2023). Exploring the use of large language models (LLMs) in chemical engineering education: Building core course problem models with Chat-GPT. Education for Chemical Engineers, 44, 71–95. https://doi.org/10.1016/j.ece.2023.05.001

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (arXiv:2201.11903). arXiv. http://arxiv.org/abs/2201.11903

Zhang, H., Xie, J., Wu, C., Cai, J., Kim, C., & Carroll, J. M. (2024). The Future of Learning: Large Language Models through the Lens of Students (arXiv:2407.12723). arXiv. http://arxiv.org/abs/2407.12723

Copyright statement