Review of: "Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education"

Birgit Popp[1]

1 Fraunhofer Institute for Algorithms and Scientific Computing SCAI

Potential competing interests: No potential competing interests to declare.

The paper “Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education” introduces LLM and chatbots and then continues to present a case study of applying LLMs as eductational tools in Graduate Engineering Education. The authors emphasize the usefulness and potential of LLMs as educational tools.

While I agree that there is potential for using LLMs as educational tools and commend the authors in sharing their experiences, I suggest that for considering scientific publication of the article the authors may focus more on describing the actual implementation (and potentially making it available for others) and expanding on the evaluation of the implementation. This would create a stronger case for their argument to apply LLMs as educational tools.

Global comments:

The paper appears well and clearly written. However after reading a few paragraphs I developed the distinct impression that a LLM might have been used to generate significant parts of the text. It is Okay to use LLMs for improving writing, however I expect the authors to critically review and adapt the text. For example, in the fourth paragraph it says: "emphasizing equity, transparency," however these are the only mentions of the words “equity” and “transparency” in the text, which suggests that in fact these concepts are not emphasised by the authors. This suggests that the terms were used arbitrarily without referring to content to be elaborated on in a later section of the paper. As I do not believe that the authors used these words arbitrarily, I suspect they might have been generated by AI.

Moreover, overly enthusiastic formulations like “revolutionizing the educational space in diverse and innovative ways.” suggest uncritical interpretation and lack of scientific rigour. This is only one of several examples of overuse of unspecific adjectives. This gives the text a feel of a blog post rather than a scientific article. Again, I do not believe that this reflects the authors style of writing, but might have been introduced by using AI.

Personally, I would suggest to the authors to either delete Sections 2 and 3 or significantly shorten those sections and refer to existing literature, e.g. https://arxiv.org/pdf/2303.11717.pdf or https://www.mdpi.com/2071-1050/15/5/4012.

What is the purpose of Section 4 “Chatbots: Strategies and Tools”? It reads like a verbose collection of example inputs and description of outputs. This is not systematic evaluation but a case-study-like narration.

The main contribution of the paper seems to be in Section 5 “Chatbot Implementation: Case Study and Evaluation of a Graduate STEM Course Incorporated with ChatGPT”, where the specifics of implementing a Chatbot for a Graduate STEM course are discussed in addition to a first step in evaluating the implementation. However, the authors admit themselves, that at this stage, Section 5 constitutes a case study. No systematic evaluation has taken place. Moreover, the authors do not share their dataset on what constitutes correct answers. Thus, their findings can not be reproduced. Moreover, I am wondering about subjective impressions of students and academic staff interacting with the chatbot and user behaviour. Did students or academic staff adopt the chatbot? How many of them? What were challenges and experiences they had?

Minor comments:

- Glossary: Not all terms are listed as new line items: AI, ML and DL are lumped together in one paragraph and it is unclear why.

- The second paragraph starting with “The dawn of the 21st century marked the rise of AI […]” emphasises positive transformation through digital technology. While there has undoubtedly been positive impact, there have been also challenges brought on through digital technology adoption in general and AI in specific. For example, both students and teaching staff suffered from the widespread adoption of video conferencing software for teaching during the Covid pandemic. More recently, teaching staff is struggling with the implications of LLMs writing essays or other marked material for students, and their inability to detect those AI written texts. Recently OpenAI announced that an internal project of theirs that was aiming to develop a high accuracy tool to detect AI written texts was discontinued, due to the low accuracy of the tool, when presented with texts written by modern LLMs. That is, OpenAI thereby suggested that texts written by LLMs may be indistinguishable from human written content, and thus text integrity can not be guaranteed with acceptable accuracy. I believe in addition to highlighting positive developments, negative effects should also be critically discussed.

- Figure 1 is inaccurate and it is unclear what purpose Figure 1 is serving. Inaccuracies in Figure 1 include: Subsummation of NLP as a step in LLM generation, next to Corpus and Deep Learning. NLP is a field of study and work that may include corpus analysis, deep learning and large language models. Another inaccuracy in Figure 1 is that the Figure suggests that next to Chatbots only three other applications of LLMs exist. That is inaccurate. Many more applications exist, for example content generation, classification (more broadly than sentiment analysis), knowledge extraction, paraphrasing, code generation, etc. Another inaccuracy is to suggest that RLHF and human alignment are necessarily part of building chatbot applications. They are not. To the best of my understanding the purpose of the figure is to highlight for readers that LLMs are distinct from the applications they are used in, in specific, they are distinct from Chatbots, which have become the most salient example of LLMs through ChatGPT. This insight can be given to readers without Figure 1, in particular as Figure 1 is inaccurate.

- LLM Architecture: The author claims that “The advent of transformer architecture [49] has revolutionized the field by

introducing attention mechanisms". This is not true. Prior architectures like recurrent neural networks and convolutional neural networks used attention mechanisms. However, these architectures had limitations in capturing long-range dependencies and were computationally expensive for large sequences.

- 2.1. fine-tuning is presented over-simplisticly. Fine tuning may include distillation, transfer learning, prompt tuning, RLAIF and others which are not captured by the description in this section.

- Section 2.2. should include sustainability, cost of training and maintenance, Difficulty to distinguish between real knowledge and convincingly written but unverified model output, and others see for example: https://www.sciencedirect.com/science/article/pii/S1041608023000195

- Figure 4 suggests that merits of chatbot implementation for student educational purposes is "consistency and accuracy in delivering" including quality assurance of the reponse and private feedback. However the authors suggest in Section 2.2. that Hallucination, Outdated Knowledge, harmful content, privacy and over reliance are _challenges_ when working with LLM. How does this square with claiming that merits of chatbot implementation include what is depicted in Figure 4? Moreover other merits are claimed (e.g. fast and efficient response and easy access) that have been identified as challenges by other authors.