OPEN CHALLENGES OF ENGLISH LANGUAGE STUDIES AND TEACHING IN THE ERA OF LARGE LANGUAGE MODELS: A BRIEF REVIEW AND DISCUSSION

Phuong Anh Do[1](MA)

Abstract

Large Language Models (LLMs) are renowned for their robust language comprehension and generation capabilities, suggesting their potential to positively influence the field of English Language Studies and Teaching. However, the integration of these models into educational settings is not without complications. This paper critically reviews the nature and capabilities of LLMs, addressing the substantial challenges associated with their adoption in English language education. Key issues include concerns over academic integrity; the potential overestimation of LLM capabilities and significant barriers related to access and inclusion. Through this exploration, the paper highlights the nascent state of LLMs in educational contexts and the urgent need for careful consideration and strategic planning before these models can be fully integrated into the English language teaching framework.

Keywords: LLMs, second-language teaching, second-language studies.

1. INTRODUCTION

Since the widespread introduction of ChatGPT in November 2022 (OpenAI, 2023), the global community has recognized the potential of Large Language Models (LLMs) to generate human-like textual content. Trained on vast amounts of data from diverse domains, LLMs have become central to numerous software tools designed to assist with a variety of complex text-based tasks, such as document analysis[3], generative search[4], proofreading[5], and software programming[6]. These capabilities have led to the proposal that LLMs could significantly benefit English Language Studies and Teaching by leveraging their ability to comprehend and generate English text (Teaching and Hub, 2023). However, this assumption overlooks critical gaps in understanding between the AI/Natural Language Processing community that develops LLM-powered softwares and the educational research community that addresses practical issues in language education. This paper aims to bridge this gap by reviewing the terminology and capabilities of LLM technology and examining the challenges that educators and students face when integrating these models into educational settings. The paper is organized as follows:

• Section 1: Introduction

• Section 2: Review of essential LLM terminologies necessary for educational researchers to understand the true nature and capabilities of LLM technology and its associated software tools.

• Section 3: Exploration of potential challenges in adopting LLMs and the softwares they powered within the English language teaching framework, including concerns about academic integrity, as LLMs could undermine traditional assessment methods; unrealistic expectations of LLM capabilities, focusing on their roles as reliable vocabulary sources and multilingual translators; and issues of access and inclusion, which could worsen existing digital divides and educational inequalities.

[1] Corresponding author: anhdp@ftu.edu.vn
Faculty of English for Specific Purposes, Foreign Trade University

[3] https://casetext.com/
[4] https://www.perplexity.ai/
[5] https://www.quillbot.com/
[6] https://github.com/features/copilot/

• Section 4: Discussion of related research works.
• Section 5: Presentation of the limitations of this research.
• Section 6: Conclusion.

2. REVIEWING TERMINOLOGIES OF LARGE LANGUAGE MODELS

- *Large Language Models:* These are sophisticated Artificial Neural Networks designed to predict the next word (token) in a sentence. Initially, without any specific adjustments, pre-trained large language models primarily perform text completion tasks by auto-regressively estimating the probability of subsequent tokens (Bowman, 2023).

- *Instructionally Aligned Language Models:* These are a subset of LLMs that undergo further fine-tuning with human-generated instructions as inputs. This process equips the models to generate text that not only completes the input but does so in a manner that aligns with demonstrated human behaviors (Brown, 2020). Prominent examples include OpenAI's GPT family (OpenAI, 2023a), Google's Gemini (Team, 2024), and Meta's Llama family (Touvron et al., 2023). In this paper, the term LLMs will refer specifically to these instructionally aligned versions.

- *LLM-powered Software:* This refers to applications developed to leverage the aligned outputs of LLMs to address specific tasks. These tasks are usually handled via direct or indirect user prompts that guide the model's response.

- *ChatGPT*[7]*:* An example of LLM-powered software, ChatGPT provides a conversational interface that allows users to prompt in dialogues with OpenAI's GPT-3.5 or GPT-4 models. This facilitates a dynamic, multi-turn conversation experience.

3. OPEN CHALLENGES

In this section, the paper will discuss open challenges regarding the applications of LLMs and LLM-powered tools in the context of English Language Study and Teaching.

3.1 Academic Integrity and Assessments:

The adoption of LLM-powered software in educational settings raises significant concerns regarding academic honesty and the integrity of assessments. Traditional methods of text-based assessment are becoming increasingly vulnerable in environments where students have access to LLM based text-completion tools. A typical 250-word essay translates to approximately 340 tokens, a fraction of the maximum processing capability of advanced models like GPT-4 Turbo and Gemini 1.5[8]. This capability poses a risk that students might exploit these tools to complete assignments with minimal personal effort, potentially undermining the assessment of their true understanding and abilities. Moreover, the ease with which LLMs can generate coherent and contextually appropriate text challenges the validity of traditional writing assignments. If language proficiency, particularly in a second language, is supposed to reflect an individual's ability to use the language for practical communication purposes (Mayo, 2000), reliance on AI-generated text could detract from the authentic development of these skills. This scenario raises re-open questions about the future role of writing tasks in language learning and assessment (Li & Li, 2022). As AI tools become more sophisticated and accessible, educators must reassess the methods they use to teach and evaluate language proficiency. This may include incorporating more interactive and practical language use assessments that are more resistant to automation, such as oral examinations, real-time discussions, and group projects. Such approaches encourage students to actively use the language in real-world situations, thereby fostering a deeper and more genuine assessment of language acquisition.

[7] https://github.com/features/copilot
[8] https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

3.2. Over-expectations of Language Models Capabilities

3.2.1. LLMs as reliable and plentiful vocabulary sources?

The Principle of Least Efforts (Fisher et al., 2005) stated that information seekers use the most convenient search method in the least exacting mode available, which is perfectly valid in the case of LLMs users. The evolving ability of large language models to generate human-like answers can mislead students into using them as a reliable source of knowledge. For instance, language learners can now achieve a detailed explanation of jargonized terminology via a simple prompt on the conversation interface from ChatGPT, instead of browsing multiple websites via conventional search engines. This behavior can increase search efficiency and satisfaction (Spatharioti et al., 2023), but it is extremely dangerous if being overused. Due to the probabilistic nature of the Transformer architecture (Vaswani et al., 2023) and the possibly out-of-date training data (Cheng et al., 2024), autoregressive language models allow the production of untruthful but convincing contents, resulting in a phenomenon called "hallucination" (Huang et al., 2023). Irresponsibly adapting output hallucinated from the LLMs regardless of contexts and verification can lead to an inappropriate and inadequate understanding of the subject matter. Search engine providers such as Google or Bing tried to mitigate this problem by designing retrieval-augmented generation systems (RAGs) (Lewis et al., 2021) which incorporate page-ranked results into the LLM prompts. Thanks to in-context learning abilities (Brown, 2020), these generative search engines can deliver fluent and informative statements with references provided, but the verifiability and accuracy of these citations are still put under questions (Liu et al., 2023). To mitigate these risks, educational institutions and users must promote information literacy and critical evaluation skills. This includes teaching how to cross-check information from multiple sources, understand the limitations of AI-generated content, and apply context-specific judgment. Encouraging the development of these competencies helps ensure that while LLMs serve as a powerful tool for vocabulary learning and information retrieval, they complement rather than replace the rigorous, discerning approach necessary for genuine understanding and innovation.

3.2.2. LLMs as multilingual translators?

The generative and comprehensive abilities of LLMs on multiple languages remain fundamental questions, given English language students come from a variety of nationalities and professionals. There is proof that LLMs express a proficiency bias toward high-resource languages, achieving better explanation in common languages such as English, and German over less popular ones (Xu et al., 2024). This bias originates from systematic inequalities of language resources in both the training data and the human evaluators (Blasi et al., 2021). Common users tend to overlook this deficiency, due to the high performance of LLMs in solving translation-equivariant tasks, such as text summarization, text-refinement or text completion (Zhang et al., 2023). However, for language learners, it is the translation-variant tasks that require the most attention, including back- and forward-translation. LLMs's underperformance in this category of tasks hinders the accurate grasp and mastery of nuanced language aspects, crucial for language learners. Tasks such as idiomatic expressions, colloquialisms, and cultural references often present significant challenges, which LLMs can misinterpret or oversimplify due to both their inherent biases and the insufficient amounts of contexts provided in the user's prompts. This underperformance can lead to misconceptions or incomplete understanding of the language for students relying on these tools for learning.

3.3 Access and Inclusion:

One of the primary challenges in integrating large language models (LLMs) into

education is providing equitable access. High-performing open-source LLMs, such as Grok[9] which contains 314 billion parameters, require extensive computing resources like multiple A100 GPUs for operation. This makes their deployment in educational environments impractical. Cloud providers offer various solutions to this issue, including renting computing resources or subscribing to cloud-hosted proprietary LLMs, such as Azure's GPTs or Vertex AI's Gemini. However, accessing these tools still necessitates a reliable internet connection. The longstanding issue of access to modern computational resources and reliable internet, a major factor in the digital divide of language and foreign language teaching (Yaman, 2015), is now exacerbating a new challenge known as the "AI divide" (Kitsara, 2022). This gap further widens the existing disparities between students from well-resourced educational institutions and those from less privileged backgrounds. Quantitative evidence shows that students with formal educational credentials—such as Bachelor's, Master's, and professional degrees—are more likely to have exposure to LLMs and LLM-powered software in the future, compared to those without such qualifications (Eloundou et al., 2023). This situation demands urgent attention from policymakers, educational leaders, and technology providers to develop more inclusive strategies. These might include subsidized access to AI technologies, grants for upgrading technological infrastructure, or even partnerships between technology companies and educational districts.

4. RELATED WORKS

Different works have been published regarding the impacts of LLMs research and the derived technology to one or more subdomains of education. For example, there are existing works on how LLMs and Generative AI as a whole can alter the way students approach introductory programming and computing courses (MacNeil et al., 2024, Denny et al., 2024), given its powerful ability to generate programming code. A recent user study investigates the effectiveness of LLM-generated feedback on upper secondary students' writing, finding that feedback from GPT-3.5-turbo improved essay revisions, task motivation, and positive emotions compared to no feedback, demonstrating the potential of LLMs for timely, impactful feedback in educational settings (Meyer et al., 2024).

Looking at a broader picture of LLMs' impacts on educational research (Wang et al., 2024) conducts a recent comprehensive review of the technologies and applications of Large Language Models (LLMs) in education, examining their benefits, datasets, challenges, and future research opportunities to enhance personalized learning and educational practices. (Wong and Looi, 2024) explore the transformative impact of Generative AI on education, focusing on learning, teaching, and assessment, with contributions from scholars in East and Southeast Asia. The Berkman Klein Center published a blog post envisioning future disruptions of LLMs on teaching, learning and discussing necessary policies needed to be implemented to enable safe uses of the technology (Ha, 2023). At a more macro level, OpenAI conducted a potential forecast on the future of labor markets where LLMs have been developed to become a general-purpose technology, emphasizing on the important role of education as a proxy for skills acquisition (Eloundou et al., 2023).

5. LIMITATIONS

This position paper provides a focused discourse on the potential complications that LLMs and their associated software could introduce into the settings of English language studies and teaching. However, it does not address issues related to curriculum adaptations or

[9] https://x.ai/blog/grok-os

pedagogical strategies that might be essential in integrating these technologies effectively. This omission stems from a limitation in the scope of the author's expertise and the insights available at the intersection of AI, linguistics, and pedagogical research. Effective integration of LLMs into language education requires a multidisciplinary approach, involving contributions from experts in AI and linguistics, as well as those in educational theory and practice.

The absence of a detailed exploration into curriculum adaptations and teaching strategies highlights a critical gap in the discussion. This gap suggests the need for further research and collaboration across these diverse fields to develop comprehensive strategies that not only mitigate the risks but also enhance the potential benefits of LLMs in language learning environments. Future work should aim to bridge these disciplinary divides, fostering a dialogue that could lead to more robust and pedagogically sound applications of language model technologies in educational settings.

6. CONCLUSION

In conclusion, while Large Language Models (LLMs) hold considerable promise for advancing English Language Studies and Teaching, their integration into educational settings presents significant challenges that must be carefully addressed, due to both the intrinsic properties of the LLMs and the Concerns over academic integrity highlight the need for new assessment methods that resist automation and better reflect students' genuine language skills. The tendency to overestimate LLM capabilities necessitates a focus on developing information literacy and critical evaluation skills among students to prevent reliance on potentially inaccurate AI-generated content. Furthermore, issues of access and inclusion underscore the importance of equitable distribution of resources to avoid exacerbating existing educational inequalities.

REFERENCES

Blasi, D., Anastasopoulos, A., & Neubig, G. (2021). *Systematic inequalities in language technology performance across the world's languages*.

Bowman, S. R. (2023). *Eight things to know about large language models*.

Brown, T. B. (2020). *Language models are few-shot learners*.

Cheng, J., Marone, M., Weller, O., Lawrie, D., Khashabi, D., & Durme, B. V. (2024). *Dated data: Tracing knowledge cutoffs in large language models.*

Denny, P., Prather, J., Becker, B. A., Finnie-Ansley, J., Hellas, A., Leinonen, J., Luxton-Reilly, A., Reeves, B. N., Santos, E. A., & Sarsa, S. (2024). Computing education in the era of generative AI. *Communication of the. ACM*, *67*(2), 56–67.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *Gpts are gpts: An early look at the labor market impact potential of large language models.*

Fisher, K. E., Erdelez, S., & McKechnie, L. (2005). *Theories of information behavior.*

Ha, Y. J. (2023). *Exploring the impacts of generative ai on the future of teaching and learning.*

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). *A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.*

Kitsara, I. (2022). *Artificial Intelligence and the Digital Divide: From an Innovation Perspective*, pages 245–265. Springer International Publishing, Cham.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küt- tler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval- augmented generation for knowledge-intensive NLP tasks.*

Li, J. & Li, M. (2022). Assessing l2 writing in the digital age: Opportunities and challenges. *Journal of Second Language Writing*, 57:100913. L2 writing assessment in the

digital age.

Liu, N. F., Zhang, T., & Liang, P. (2023). *Evaluating verifiability in generative search engines*.

MacNeil, S., Leinonen, J., Denny, P., Kiesler, N., Hellas, A., Prather, J., Becker, B. A., Wermelinger, M., & Reid, K. (2024). Discussing the changing landscape of generative ai in computing education. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*, SIGCSE 2024, page 1916, New York, NY, USA. Association for Computing Machinery.

Mayo, M. d. P. G. (2000). Focus on form in classroom second language acquisition. Catherine Doughty and Jessica Williams (eds.). New York: Cambridge university press, 1998. pp. xiv + 301. 64.95*cloth*,22.95 paper. *Studies in Second Language Acquisition*, *22*(1):123–124.

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.

(OpenAI, 2023a) OpenAI (2023a). *Gpt-4 technical report.*

(OpenAI, 2023b) OpenAI (2023b). *Introducing chatgpt.*

Spatharioti, S. E., Rothschild, D. M., Goldstein, D. G., & Hofman, J. M. (2023). Comparing traditional and LLM-based search for consumer choice: A randomized experiment.

Teaching, J. & Hub, L. (2023). *Benefits of LLMs in education*.

Team, G. (2024). *Gemini: A family of highly capable multimodal models.*

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *Llama: Open and efficient foundation language models*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention is all you need.*

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., & Wen, Q. (2024). *Large language models for education: A survey and outlook.*

Wong, L.-H. & Looi, C.-K. (2024). Advancing the generative AI in education research agenda: Insights from the Asia-pacific region. *Asia Pacific Journal of Education*, *44*(1), 1–7.

Xu, S., Dong, W., Guo, Z., Wu, X., & Xiong, D. (2024). *Exploring multilingual concepts of human value in large language models: Is value alignment consistent, transferable and controllable across languages?*

Yaman. (2015). Digital divide within the context of language and foreign language teaching. *Procedia - Social and Behavioral Sciences*, *176*, 766–771.

Zhang, X., Li, S., Hauer, B., Shi, N., & Kondrak, G. (2023). Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.