On the application of Large Language Models for language teaching and assessment technology

Andrew Caines\textsuperscript{1}, Luca Benedetto\textsuperscript{1}, Shiva Taslimipoor\textsuperscript{1}, Christopher Davis\textsuperscript{1}, Yuan Gao\textsuperscript{1}, Øistein Andersen\textsuperscript{1}, Zheng Yuan\textsuperscript{2,1}, Mark Elliott\textsuperscript{3,1}, Russell Moore\textsuperscript{1}, Christopher Bryant\textsuperscript{4,1}, Marek Rei\textsuperscript{5,1}, Helen Yannakoudakis\textsuperscript{2,1}, Andrew Mullooly\textsuperscript{3}, Diane Nicholls\textsuperscript{6} and Paula Buttery\textsuperscript{1}

\textsuperscript{1}ALTA Institute & Computer Laboratory, University of Cambridge
\textsuperscript{2}King’s College London
\textsuperscript{3}Cambridge University Press & Assessment
\textsuperscript{4}Writer, Inc.
\textsuperscript{5}Imperial College London
\textsuperscript{6}English Language iTutoring (ELiT)

Abstract
The recent release of very large language models such as PaLM and GPT-4 has made an unprecedented impact in the popular media and public consciousness, giving rise to a mixture of excitement and fear as to their capabilities and potential uses, and shining a light on natural language processing research which had not previously received so much attention. The developments offer great promise for education technology, and in this paper we look specifically at the potential for incorporating large language models in AI-driven language teaching and assessment systems. We consider several research areas – content creation and calibration, assessment and feedback – and also discuss the risks and ethical considerations surrounding generative AI in education technology for language learners. Overall we find that larger language models offer improvements over previous models in text generation, opening up routes toward content generation which had not previously been plausible. For text generation they must be prompted carefully and their outputs may need to be reshaped before they are ready for use. For automated grading and grammatical error correction, tasks whose progress is checked on well-known benchmarks, early investigations indicate that large language models on their own do not improve on state-of-the-art results according to standard evaluation metrics. For grading it appears that linguistic features established in the literature should still be used for best performance, and for error correction it may be that the models can offer alternative feedback styles which are not measured sensitively with existing methods. In all cases, there is work to be done to experiment with the inclusion of large language models in education technology for language learners, in order to properly understand and report on their capacities and limitations, and to ensure that foreseeable risks such as misinformation and harmful bias are mitigated.

Keywords
large language models, education technology, natural language processing, question difficulty estimation, text generation, automated assessment, grammatical error correction, responsible AI
1. Introduction

The training of large language models (LLMs) – also known as pre-trained language models or foundation models – has had a transformative effect on the fields of natural language processing (NLP) and artificial intelligence (AI) more broadly. LLMs are ‘large’ because they are neural networks made up of billions or trillions of parameters. The networks are Transformers [1] trained on huge swathes of text from the World Wide Web, using language modelling objectives such as predicting omitted (or, ‘masked’) words and sentence pairs [2], or predicting the next token in a sequence [3]. Furthermore, in the few-shot learning paradigm, LLMs can be directed towards new tasks without large quantities of task-specific data [4, 5], the collection of which tends to be time-consuming and costly. Overall, LLMs also offer great potential for educational applications. One previous paper has already provided an overview of some of the possible applications of LLMs to educational technology as a whole [6], across subjects. Our distinct contribution is to focus on the language learning and assessment domain specifically. In this paper, we describe some of the uses for LLMs in the context of language learning, discuss the state of the art or work in progress, and consider practical, societal and ethical implications.

We set out a number of uses for LLMs in the language learning domain, relating to content creation and calibration, automated assessment of written texts, and personalised feedback. In each case the general principles of the approach are well established thanks to previous work with pre-existing LLMs such as BERT [2] – language models with millions of parameters have existed for several years already. We look at the opportunities presented by the recent and rapid steps taken by OpenAI in releasing new variants from the ‘generative pre-training’ (GPT) model series, along with some newly published pre-prints relating to LLMs and the language learning research field. We refer to some LLM-driven language learning applications already in use, and outline the variety of LLMs available besides GPT. It is a fast evolving research field, one being driven by industry developments. We perceive some possible risks in this research trajectory, which include but are not limited to the absence of proper safeguards on education technology, the lack of public understanding as to how LLMs are trained and how they can confidently assert incorrect information, and the harm to the advancement of education technology as a whole if it is considered ‘solved’ by investors and research councils – not to mention the ethical issues that are already well known, such as data protection [7], examination malpractice [8, 9], environmental impact [11, 12], and internet addiction [13, 14, 15], among others.

2. Large Language Models & Language Learning EdTech

At the time of writing, one of the most prominent LLMs is OpenAI’s GPT-4 [16] released in March 2023 after six months of pre-launch work improving safety and fact-checking, also for product development with selected partners including the education technology (EdTech) firms Duolingo² and Khan Academy³. This built on the prior success and notoriety of GPT-3, released

---

¹Note that the text-matching tool Turnitin [10], which is commonly used to detect plagiarism, has developed a module to detect the use of AI in essays: https://www.turnitin.com/blog/the-launch-of-turnitins-ai-writing-detector-and-the-road-ahead
²https://blog.duolingo.com/duolingo-max/
³https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/
in June 2020, along with its related chatbot application, ChatGPT\textsuperscript{4}. Recently, we have seen more focus on the effects and implications of using chatbots for creating interactions with language learners [17, 18].

Of most relevance here is the partnership between OpenAI and Duolingo, the language learning application developer, which resulted in the subscription service Duolingo Max [19]. Duolingo Max presents two new features: ‘Role Play’ and ‘Explain My Answer’. The former involves some limited conversation towards a goal such as ordering food, which starts with a pre-scripted prompt but then proceeds over several open-ended chat turns between user and chatbot. The latter is an option for additional feedback on grammatical points, involving a limited dialogue with pre-specified responses for the user to guide the conversation (e.g. “Yes, I’m all set”, “Can I see an example?”, “No, please elaborate”). Another limitation is that the service is only currently available in selected countries and for a few languages.

Nevertheless, this development points towards further opportunities in AI-driven education technology for language learning, as discussed below. It should be noted that there are many alternatives to the GPT models, including the ‘text-to-text Transformer’ (TS) [20], PaLM (Parallel Language Model) [21] and LaMDA (Language Model for Dialogue Applications) [22] by Google; LLaMA [23] and Open Pre-trained Transformers (OPT) [24] by Meta AI; and DeepMind’s Gopher [25]. At least a few of these models are ‘multilingual’, having been trained on corpora from multiple languages, albeit with a strong bias towards English\textsuperscript{5}. In addition there are models which have been trained on bilingual data, notably Chinese–English [26] and Russian–English\textsuperscript{6}. Alongside LLM developments by large technology companies, we also note the various open-source efforts to train on known datasets (e.g. EleutherAI’s GPT-X [27] and The Pile [28]), or as massive research collaborations (e.g. BLOOM: the BigScience Large Open-science Open-access Multilingual Language Model [29]), or to democratise LLMs for wider use in web applications involving natural language interfaces (e.g. langchain\textsuperscript{7}, Cohere AI\textsuperscript{8}, and Transformers [30]). There are also open-source alternatives to ChatGPT, such as Open Assistant\textsuperscript{9} and StableVicuna\textsuperscript{10}.

Finally we highlight efforts to transparently evaluate LLMs in comprehensive and varied ways, for instance in the HELM project (Holistic Evaluation of Language Models) [31].

3. **Content Creation: Creating Assessment Items and Teaching Materials**

Pre-trained transformer models are being explored to generate exam texts and items for educational purposes like in the Duolingo English Test [32], or in Google’s quantitative reasoning application, Minerva [33]. Variations of GPT models are best known to the wider public as being able to create common forms of language assessment tests. However, there are other successful

\footnotesize{\textsuperscript{4}https://openai.com/blog/chatgpt \\
\textsuperscript{5}e.g. See the distribution of languages in the training data for GPT-3: https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_document_count.csv \\
\textsuperscript{6}https://github.com/yandex/YaLM-100B \\
\textsuperscript{7}https://python.langchain.com/en/latest/index.html \\
\textsuperscript{8}https://cohere.com/ \\
\textsuperscript{9}https://open-assistant.io/ \\
\textsuperscript{10}https://stability.ai/blog/stablevicuna-open-source-rlhf-chatbot}
pre-trained LLMs such as variations of BERT [2], BART [34], or T5 [20] which are popular among NLP scientists. Since these models are trained with different target task objectives, they should be prompted differently for various kinds of text generation. None of them has been trained with a storytelling objective to generate fluent long texts as GPT* models have. Nevertheless, since the target tasks are better defined, evaluation of the performance of these models is more thorough and explainable. BERT-based models have shown impressive results in filling the gaps in sequences of text. BART achieves state-of-the-art results in generating parallel sentences as in machine translation or grammatical error correction [34, 35]. T5 modelled 24 tasks as text-to-text generation and proved very successful in question answering and text summarisation [20].

These models are widely used for narrower tasks like question generation [36], for reading comprehension exercises [37], or prompt generation for writing and speaking. For example, Felice et al. [38] use BERT and ELECTRA [39] to predict the position of the gaps for designing high-quality cloze tests for language learners, while various pre-trained language models are used for generating distractors [40, 41, 42]. In addition LLMs have been put to use for the purpose of text simplification, which is relevant for language learners in the context of reading comprehension exercises and adapting texts automatically to an appropriate level. Notably, a GPT-3 based solution by the UniHD team was the winning entry for the English track of the TSAR-2022 Shared Task on Multilingual Lexical Simplification [43, 44].

Datasets and Evaluation With the emergence of LLMs, large-scale datasets are required for evaluation. Text generation methods are evaluated using text-similarity based metrics like BLEU [45], ROUGE [46], METEOR [47], and more recently BERTScore [48], or learned evaluation metrics [49, 50] which assess the correlation between generated texts (e.g. the question) and the ones originally written by human experts. Automatic evaluations require top-quality and expert-designed datasets. Available datasets for language learning exams for NLP research include RACE [51], SCDE [52] and CLOTH [53]. However, there are smaller-scale datasets such as CEPOC [54] for Cloze test creation, and the Teacher-Student Chatroom Corpus [55], which can be used as test sets to evaluate zero-shot or few-shot learning models. Evaluation approaches for open-ended text generation are still far from being ideal. Human-centric evaluations involve ranking the generated texts based on different factors, such as fluency and coherence of the generated texts, or its relevance to the context document and the answer (where available) [56]. In this new era of increasingly large language models, human evaluation is more difficult and time-consuming, leading researchers to design comparison datasets that contain human-labelled comparisons between outputs of different systems [57].

Human-in-the-loop content generation As an exploratory study, we have worked with publicly available GPT-3 models to generate open-ended texts and evaluate their suitability as a basis for low-stakes, self-study language learning exercises. Having a human-in-the-loop policy in mind, the prompts are engineered by a human expert, with post-generation text refinement and question authoring also carried out by experts. Such an approach can be seen as helping mitigate against various known risks associated with the output of LLMs (e.g., hallucinations, offensive content, stereotyping, etc).
All definable model parameters (Temperature, Frequency Penalty, Presence Penalty and Max Length) are kept to fixed levels throughout to limit the number of variables across the dataset. Input prompts containing target genre and key content points are designed by the human expert in order to provide a basis for possible testing foci at the target level. The key content points also help generate similar enough output texts from a single (or slightly modified) input prompt, to allow for collation of the best elements from multiple output versions.

For this research, the generated texts are intended to support single B2 CEFR level\(^\text{11}\) multiple choice reading comprehension questions with 3 answer options. The generated texts are reviewed by the human expert and given an ‘accept’ or ‘reject’ status based on their appropriateness for the target proficiency level and relevance to the content points. Accepted texts are added to a content pool, also containing fully human-authored texts. Another group of human experts (question writers) approach the accepted texts as they would any other content. For openness and transparency, question writers are informed in advance that the pool contains AI-generated content, but not which texts are AI-generated, and which have been written by human authors. Question writers select and edit the texts, writing one 3-option multiple choice question per text. The annotations of this dataset, including the accept/reject status and the measures of quality of the generated texts assessed by question writers (based on the amounts of edits made on the texts) can be used to train models which can automatically assess the generated texts in future. The dataset can also be used to train reward functions for further fine-tuning of the generative models.

ChatGPT has been trained using a combination of supervised fine-tuning and reinforcement learning from human feedback (RLHF) [58, 59]. It uses InstructGPT [57] and includes the steps: pre-training, fine-tuning, and reward learning. The reward function used to fine-tune InstructGPT is trained using a dataset of pairs of generated texts with a human-labelled judgement on which text is better, with the objective to maximise the score difference between the ‘winning’ and ‘losing’ texts. The purpose of collecting coarse human labelling is to mitigate any mismatch between the true objective and the preferences of human annotators, thus increasing inter-annotator agreement [59]. Nevertheless, relying solely on general annotations, as in the case of InstructGPT, results in a reward function that fails to shed light on the quality of texts across various aspects, making it too broad to apply in narrower tasks and fields. To address this limitation, we can take advantage of the existing high-quality annotations available to us from skilled and experienced professional human annotators. By exploiting their expertise, we can train a more nuanced reward function that offers fine-grained evaluation and provides interpretable scores, aligning more effectively with our specific research goals. Finally, we can evaluate different methods of content generation on our reading practice platform, Read&Improve\(^\text{12}\) [60]. By collecting both implicit user feedback – which texts they engage with more by spending longer reading them, clicking on definitions, completing the tasks, etc – and explicit user feedback (e.g. by asking them to rate texts and express opinions) we can assess which LLM-driven systems are most successful.

\(^\text{11}\)The Common European Framework of Reference for Languages (CEFR) organises language proficiency in six levels, A1 to C2, which represent Basic User, Independent User and Proficient User.

\(^\text{12}\)https://readandimprove.englishlanguageitutoring.com/
4. Calibrating Assessment Items and Teaching Materials

In addition to content creation, LLMs can potentially be leveraged for the evaluation and calibration of existing learning content: test items and teaching content. An example of this is question difficulty estimation (QDE) from text, which has received increasing research interest in recent years [61, 62]. QDE from text offers a way to overcome the limitations of traditional approaches such as manual calibration and statistical analysis from pre-testing. These traditional approaches are either subjective or introduce a long delay between item creation and deployment due to the complexities of pre-testing on sizeable and representative populations.

QDE from text is a regression task, where the model is asked to provide, given the text of the question, a numerical estimation of its difficulty on a given scale. It can be either a supervised or unsupervised task, depending on whether a dataset of already calibrated exam questions is available, and LLMs have been used in both scenarios. Regarding supervised estimation, LLMs that leverage transfer learning – specifically, BERT [2] and DistilBERT [63] – are the current state of the art [64, 65] and have been shown to outperform other approaches using traditional NLP-derived features [66]. There has been less work on unsupervised estimation but, even in this scenario, LLMs have been shown to be helpful for estimating question difficulty from text [67].

All the models proposed in previous research require some kind of transfer learning starting from the publicly available pre-trained models, which might be expensive and not feasible in all scenarios. Bigger LLMs, such as the aforementioned GPT models, could be used for zero-shot or few-shot difficulty estimation from text, which is yet to be explored. As an example, ChatGPT can be asked to rank given questions by difficulty, and it can also provide an indication of the specific difficulty level (e.g. Easy, Medium, Hard). Crucially, the difficulty of a pool of questions depends on the specific student population that is assessed with them, and it is difficult to provide the LLM with all the information required to describe the specific pool of learners that will be assessed with the items. The model seems to be – at least partially – capable of distinguishing between different CEFR levels, since the same question can be assigned different levels depending on whether the model is asked to consider learners of level A1 or C1. However, extensive experiments should be carried out to better evaluate this, as the model sometimes performs counterintuitive estimations: in our preliminary experiments, for instance, ChatGPT sometimes estimated a question to be more difficult for C1-level learners (i.e., “Proficient”) than A1-level learners (i.e., “Basic”).

5. Automated Assessment of Language Learners

Automated assessment has long been a prominent task in educational applications research: for instance assessing learner English speech [68, 69, 70, 71] and writing [72, 73, 74, 75, 76]. Here we focus on writing and the task of ‘automated essay scoring’ (AES). Whereas previous systems have involved feature engineering – typically centred around informative sequences of characters, words, part-of-speech tags, as well as phrase structures from a parser and automatically detected errors [73, 77] – more recent research systems have involved neural models for assessment [74, 75, 76]. These models tend to be carefully crafted and evaluated, since language assessment
can be a task with major consequences for the learner, including education and career prospects. Therefore any involvement of LLMs in assessment systems must be approached cautiously and its impact measured on existing benchmarks. Deployment of LLM-based assessment models should be restricted to human-in-the-loop low-stakes contexts first, including practice applications [77, 60] or placement tests such as Linguaskill[13].

The idea of using ChatGPT for assessing students’ answers was put forward by Jeon & Lee [78]. Further practical steps were taken by Mizumoto & Eguchi [79] who experimented with GPT-3.5 for AES on 12,000 essays from the ETS Corpus of Non-Native Written English (TOEFL11) [80], compared the scores to benchmark levels on a 0-9 scale, and concluded that a GPT-only model only achieves weak agreement with the reference scores (.388 quadratic weighted kappa). The authors furthermore compare the GPT scorer with several models involving various combinations of 45 linguistic features – related to lexical diversity, lexical sophistication, syntactic complexity and dependency, and cohesion – and observe that although the GPT baseline is outperformed by the linguistic features on their own, the best results are obtained by combining the two approaches (.605 QWK). This is a finding similar to previous research, as the previous state-of-the-art performance was obtained by combining BERT-style neural models and feature-based models [76, 81].

One potential use for LLMs regarding assessment that, to the best of our knowledge, has not been thoroughly explored is for explaining assessment predictions. Explainable AI is an emerging research topic, a regulatory prospect [82], and a challenge for NLP models dependent on ‘black box’ neural networks. One possibility is to adopt the ‘chain-of-thought’ prompting style [83] – as opposed to zero-shot or few-shot prompting [5] – to elicit explanations about assessment decisions from LLMs. Exactly how to engineer a series of prompts for the LLM in the chain-of-thought style is a matter for investigation, but for instance they could be similar to the following (albeit longer to elicit better explanations):

A class of students has been given the essay prompt: <essay_prompt>.
This student essay -- <example_text>
-- was given a score of <example_score>.
Explanation: the use of language and grammatical resource are advanced but there is a spelling error in the first sentence and a grammatical error in the final sentence.
This student essay -- <target_text>
-- has been given a score of <predicted_score>.
Please give an explanation why the text was given this score.

The aim of such an approach would be to obtain explanations specific to the essay, pinpointing relevant sections from the text if possible, and grounded in marking criteria for the learner’s target level so that the explanation is relevant and useful. In common with other tasks described in this paper, further research with LLMs and proper evaluation on existing benchmarks and by human experts is needed before we can definitively conclude that this is a research avenue worth exploring.

Finally we note that there are concerns around LLMs being used in fraudulent ways by learners, but that plagiarism concerns are long-standing in computer-based exam settings. If proctoring software is set up to prevent text import from elsewhere (e.g. disabling copy-and-paste keyboard shortcuts) or to detect bursty text insertion through keystroke logging, then this is one defence against exam malpractice from LLM text generation or any other online source. In this way, LLMs are an extension of a threat we are already familiar with. Furthermore, automatic detection of LLM-generated text is the subject of the AuTexTification (Automated Text Identification), part of IberLEF 2023 (the 5th Workshop on Iberian Languages Evaluation Forum) co-located with the SEPLN Conference this year. It appears that the level of performance from submitted systems has been high, outdoing a logistic regression baseline in many cases, with system descriptions to be presented in September. It may be that such systems can be employed as an additional line of defence against exam malpractice involving LLM-generated text.

6. Providing Feedback to Language Learners

As a precursor to providing automatic lexico-syntactic feedback to language learners, one requirement is to first carry out the NLP task of grammatical error detection (GED) or grammatical error correction (GEC). These tasks have a rich pedigree involving various benchmark corpora, research papers and shared tasks – all of which enable us to establish that the current state-of-the-art approach to GED and GEC tends to involve supervised fine-tuning of neural network language models using carefully annotated training data. The recent emergence of LLMs, however, offers the prospect of developing GED or GEC models which are largely unsupervised other than some examples for few-shot learning.

The challenge ahead, in common with the application of LLMs to other tasks described in this paper, is to properly benchmark LLM-based models for GED and GEC on existing corpora, so that their performance can be compared to previous models. Some preliminary work has been done towards this aim, as described in a recent survey of GEC. For instance, Wu et al. and Coyne & Sakaguchi present preliminary results applying LLMs to GEC. The former compares ChatGPT to Grammarly and GECToR, a previous state-of-the-art GEC system, and the latter compares GPT-3.5 to two other GEC systems. Both approaches find the GPT* models perform worse than existing systems when measured using automatic evaluation techniques on existing benchmark corpora (namely CoNLL-2014, JFLEG, BEA-2019). The authors ascribe this to the model’s tendency to over-correct learner text; by inserting additional text or re-structuring phrases, the corrected text moves further from the original text and is penalised by the automatic scorers. However, both works carry out human evaluation to rate the output from each system and find a preference for the GPT* output because the corrected sentences tend to be more fluent. At the same time, they

14https://sites.google.com/view/autextification
15Note that LLM training is often described as ‘self-supervised’ due to the human-authored training data, but for the purpose of GED/GEC, we say ‘unsupervised’ because in this context no task-specific training data is required.
16Specifically, they use text-davinci-003.
found instances of under-correct in the human-generated reference sentence: in other words GPT* models catching and correcting errors which were not corrected by the expert annotators. While both approaches are preliminary and human evaluation tentative – based on only small samples of 100 sentences at a time from each test set – overly fluent corrections present a challenge for automatic evaluation methods as they are much more open-ended than minimal edits targeting grammatical errors rather than stylistic choices. Furthermore, while fluent corrections may at times be preferred by human evaluators, they may not aid language learners if they drift too far from the original text. Existing annotation guides for error correction state that edits should be as minimal as possible so that the learner can be helped to express what they are trying to say, rather than told how to express it differently: that is, how to amend an error rather than avoid it [103]. The issue is not a new one [104] but remains a matter for further investigation under the new conditions presented by more capable LLMs.

Another potential use of LLMs in this area is providing automatically-generated feedback comments to learners to explain linguistic concepts, grammatical points or semantic nuance. Indeed there was a recent shared task on feedback comment generation [105] where, when presented with an erroneous sentence such as, “He agrees the opinion”, the task was to produce a comment such as: The verb agree is an intransitive verb and cannot take direct objects: add the appropriate preposition\(^\text{17}\). Participants in the shared task were able to outperform the baseline system (an encoder-decoder with a copy mechanism based on a pointer generator network\(^\text{18}\)) through careful feature extraction from parsers and GEC models, combined with prominent LLMs at the time such as T5 or GPT-Neo [106] (e.g. Babakov et al. [107] achieved second place in the shared task; developers of the first-placed entry have not published a system description to the best of our knowledge). It remains to be seen whether current LLMs can be tuned towards even better performance on this task: it may be that the pre-processing of texts to obtain additional linguistic information and the incorporation of pre-defined templates will continue to be vital for accurate and sensible feedback comment generation, even with ever-larger LLMs involved [108]. These are methods we can trial through A/B testing of different feedback models on our essay-writing practice platform, Write&Improve\(^\text{19}\) [73, 77].

Other applications of LLMs for language learning feedback include chatbot interaction to explain linguistic concepts – akin to the ‘Explain My Answer’ feature in Duolingo Max, but also going beyond this with dialogue which is adaptive to the learner level [18] – word suggestion, paraphrasing and translation to aid learners with essay writing, and document-level feedback on, for instance, inter-sentence coherence markers, co-reference and anaphoric reference, maintaining tense and aspect consistently, argumentation structure, task completion and more. Key desiderata are that the feedback should be accurate, based on evidence, personalised, inoffensive and preferably linked to teaching materials so that the learner may continue to benefit from EdTech applications for language.

\(^{17}\)https://fcg.sharedtask.org/
\(^{18}\)https://github.com/k-hanawa/fcg_genchal2022_baseline
\(^{19}\)https://writeandimprove.com/
7. Risks & Ethical Considerations

We advocate for a cautious approach to the incorporation of LLMs in EdTech for language learning, in which the training process, performance and limitations, and pathway to delivery are well documented and the risks of misapplication of such technology are understood. There are general concerns about AI for NLP and education which are recorded in the literature and continue to be relevant, perhaps more so, as LLMs come to the fore. Firstly there is a bias towards English, and specific genres of English, due to a combination of commercial pressures, training data availability, and data sourcing from the World Wide Web: even though several models have been trained in multilingual ways, the general trend with LLMs has exacerbated this pre-existing bias [109, 110, 111]. As LLMs grow, so does their climate impact: an issue which interacts with societal and infrastructure complexities but which we should nevertheless bear in mind and attempt to mitigate [11, 12]. In addition, LLMs are known to exhibit certain biases [112] – both representational (language use around demographic groups) [113, 114] and allocational (how a system distributes resources or opportunities) [115, 113] – which need to be debiased or otherwise controlled [116, 117].

Suresh & Guttag identified various sources of harm in the ‘machine learning life cycle’ [115]: historical bias, representation bias, measurement bias, learning bias, aggregation bias, evaluation bias, deployment bias. They note that effects cascade downstream and cycle around ML systems. They provide some mitigation strategies and reference previous work in this area [118, 119]. Kasneci et al. [6] also point to copyright issues with output from LLMs which are largely unresolved, as well as concerns about pedagogical and learning effects: namely that both teachers and students, “may rely too heavily on the model”, and that it may be difficult to, “distinguish model-generated from student-generated answers” [120, 121, 122]. In addition they raise data privacy and security issues which require firmer regulation and auditing of EdTech firms, the problem of false information issued by LLMs, and of designing appropriate application interfaces which are both engaging and beneficial to end-users. It is worth noting that NLP researchers have made some attempts at using LLMs to assess the trustworthiness of generated texts, which could go some way towards mitigating the false information problem [123, 124, 125].

Regarding AIED and language learning, LLMs present specific risks relating to generated outputs which may be inaccurate, confusing, offensive, and so on – risks which are present in human teachers too, but made no less harmful as a result. For this reason the most successful systems may be human-machine hybrids, with humans in-the-loop or similar, where LLMs are viewed as assistive technology for human experts rather than replacements for them – performing the more mundane and mechanical tasks while experts provide the inputs characteristic of human interaction [126]. Another way that humans can monitor LLM outputs is through evaluation, and feedback mechanisms for systems in production, so that problematic outputs may be flagged.

We can also look at standards for ‘responsible AI’ published by technology firms and research institutes [127, 128, 129]²⁰. For example, Duolingo [130] sets out its approach to responsible AI under Validity & Reliability, Fairness, Privacy & Security, Accountability & Transparency

²⁰https://huggingface.co/blog/ethical-charter-multimodal
– all of which have been touched on in this paper. Regarding the last attribute in particular – Transparency – it is apparent from recent media stories that more can be done in this area in terms of educating the general public about how LLMs are trained, how trustworthy they may or may not be, and how best to interact with them. This is a general problem but one which nonetheless presents a challenge for EdTech applications.

8. Conclusion

In this paper, we have explored the opportunities for language-learning EdTech offered by ‘generative AI’ through LLMs. We conclude that preliminary indications are promising, but that the best systems may still require human intervention and/or the inclusion of well-established linguistic features. It may well be that LLMs can enhance language-learning EdTech, if we can establish the following through further empirical work:

1. that models enhanced by LLMs perform better than existing models on established benchmarks, or on alternative evaluation metrics which need to be defined in order to properly probe LLM capabilities for language teaching and assessment [131] – moreover that performance is sufficiently better to justify the additional costs in computing and environmental terms;
2. that LLM-enhanced technology is of benefit to language learners, whether that is measured through engagement, enjoyment, learning outcomes or some combination of the three;
3. that LLM-enhanced technology does not disadvantage relevant groups (learners, teachers, writers and editors of materials, examiners) whether through bias, misinformation, or adversely affecting student progress – instead, the technology should be assistive to all groups in some regard.

Finally, we note that LLMs should not be over-hyped as an AI revolution, but rather as an evolutionary step in neural network models – the inevitable result of the inexorable growth in network size since the Transformer was first applied to language tasks in 2017 [1]. LLMs represent a milestone on an evolutionary path which has been unfolding for many years and thus is well documented in open access publications and open source code repositories. If we maintain this tradition – by close inspection of proprietary models, or opting to use models trained in open ways – it will be of benefit both to future researchers and scientific development, but also users of AI applications who require some transparency regarding the technology. Harmful bias and other risks remain an ongoing challenge for developers of AI systems, and LLMs deployed in language learning EdTech may only exacerbate these. Therefore, proper mitigations should be put in place to address the issues which have been identified in this paper and elsewhere.

Nevertheless, LLMs present a great opportunity to continue improving EdTech for language learning, including novel ways to generate content, provide feedback, and deal with other linguistic features which hitherto have not been commonly attempted: for instance, chatting in open-ended ways at the level of the learner [18], providing document-level assessment and feedback [132], handling code-switching or ‘plurilingual’ learning [133].
Acknowledgments

This work was supported by Cambridge University Press & Assessment. We thank Dr Nick Saville and Professor Michael McCarthy for their support. We are grateful to the anonymous reviewers for their helpful comments.

References


[51] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, RACE: Large-scale ReAding comprehension


[64] Y. Zhou, C. Tao, Multi-task bert for problem difficulty prediction, in: 2020 International


[77] H. Yannakoudakis, Ø. E. Andersen, A. Geranpayeh, T. Briscoe, D. Nicholls, Developing


