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

Author 1[1,2,*,†], Author 2[3,†] and Author 3[4,†]

[1]affiliation 1

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 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, with 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 been around 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) family of models, 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, environmental impact [8, 9], and internet addiction [10, 11, 12], 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 [13] 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 Duolingo and Khan Academy. This built on the prior success and notoriety of GPT-3, released in June 2020, along with its related chatbot application, ChatGPT[1]. Recently, we have seen more focus on the effects and implications of using chatbots for creating interactions with language learners [14, 15].

Of most relevance here is the partnership between OpenAI and Duolingo, the language learning application developer, which resulted in the subscription service Duolingo Max [16]. 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,

---

[1]https://openai.com/blog/chatgpt

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 the way 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' (T5) [17], PaLM (Parallel Language Model) [18] and LaMDA (Language Model for Dialogue Applications) [19] by Google, LLaMA [20] and Open Pre-trained Transformers (OPT) [21] by Meta AI, and DeepMind's Gopher [22]. At least a few of these models are 'multilingual', having been trained on corpora from multiple languages, albeit with a strong bias towards English[2]. In addition there are models which have been trained on bilingual data, notably Chinese–English [23] and Russian–English[3]. Alongside LLM developments by large technology companies, we also note various open-source efforts to train on known datasets (e.g. EleutherAI's GPT-X [24] and *The Pile* [25]), or as massive research collaborations (e.g. BLOOM: the BigScience Large Open-science Open-access Multilingual Language Model [26]), or to democratise LLMs for wider use in web applications involving natural language interfaces (e.g. `langchain`[4], Cohere AI[5], and Transformers [27]). There are also open-source alternatives to ChatGPT, such as Open Assistant[6] and StableVicuna[7]. Finally we highlight efforts to transparently evaluate LLMs in comprehensive and varied ways, for instance in the HELM project (Holistic Evaluation of Language Models) [28].

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 [29], or in Google's quantitative reasoning application Minerva [30]. While variations of GPT models are best known to the wider public, there are other successful pre-trained LLMs such as variations of BERT [2], BART [31], or T5 [17] 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. However, 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 has 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. T5 modelled 24 tasks as text-to-text generation and proved very successful in question answering and text summarisation.

---

[2]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

[3]https://github.com/yandex/YaLM-100B

[4]https://python.langchain.com/en/latest/index.html

[5]https://cohere.com/

[6]https://open-assistant.io/

[7]https://stability.ai/blog/stablevicuna-open-source-rlhf-chatbot

These models are widely used for narrower tasks like question generation [32], for reading comprehension exercises [33], or prompt generation for writing and speaking. For example, Felice *et al.* [34] use BERT and ELECTRA [35] to predict the position of the gaps for designing high-quality cloze test for language learners, while various pre-trained language models are used for generating distractors [36, 37, 38].

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 [39], ROUGE [40], METEOR [41], and more recently BERTScore [42], or learned evaluation metrics [43, 44] 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 [45], SCDE [46] and CLOTH [47]. However, there are smaller-scale datasets such as CEPOC [48] for Cloze test creation, and the Teacher-Student Chatroom Corpus [49], 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) [50]. 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 [51].

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 level 3-option multiple choice reading comprehension questions. The generated texts are reviewed by the human expert and are 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. When released, the dataset can also be used to train reward functions for further fine-tuning of the generative models.

Following the success of ChatGPT, which uses InstructGPT [51] and benefits from reinforcement learning from human feedback (RLHF) [52, 53], we propose to use reinforcement learning to generate targeted output texts more successfully. In the development of ChatGPT, this includes the steps: pre-training, fine-tuning, and reward learning. For this purpose we rely on the pre-trained GPT models, and use our compiled dataset of generated and annotated texts (as explained above) for further reward learning and fine-tuning. 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 [53]. 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.

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 [54, 55]. 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 [56] – are the current state of the art [57, 58] and have been shown to outperform other approaches using traditional NLP-derived features [59]. 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 [60].

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 needed 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 experiments, for instance, ChatGPT sometimes estimated a question to be more difficult for C1-level than A1-level learners.

5. Automated Assessment of Language Learners

Automated assessment has long been a prominent task in educational applications research: for instance assessing learner English speech [61, 62, 63, 64] and writing [65, 66, 67, 68, 69]. 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 [66, 70] – more recent research systems have involved neural models for assessment [67, 68, 69]. 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 cautiously approached and its impacts 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 [70, 71] or placement tests such as Linguaskill[8].

The opportunity of using ChatGPT for assessing students' answers was discussed by Jeon & Lee [72]. A further step is taken by Mizumoto & Eguchi [73] by experimenting with GPT-3.5 for AES on 12,000 essays from the ETS Corpus of Non-Native Written English (TOEFL11) [74] and comparing the scores to benchmark levels on a 0-9 scale. The authors experiment with GPT-3.5, and observe that a GPT-only model achieves only weak agreement with the reference scores (.388 quadratic weighted kappa). The authors 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 [69, 75].

One potential use for LLMs regarding assessment, that to date has not been thoroughly

---

[8]https://www.cambridgeenglish.org/exams-and-tests/linguaskill/

explored to the best of our knowledge, is for explaining assessment predictions. *Explainable AI* is an emerging research topic, a regulatory prospect [76], and a challenge for NLP models dependent on 'black box' neural networks. One possibility is to adopt the 'chain-of-thought' prompting style [77] – 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 was given a score of <example_score>: <example_text>.
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 has been given a score of <predicted_score>: <learner_text>.
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.

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 [78, 79, 80, 81], research papers [82, 83, 84, 85, 86] and shared tasks [87, 88, 89] – 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 [90]. 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[9].

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 [90]. For instance, Wu *et al.* [91] and Coyne & Sakaguchi [92] present preliminary results applying LLMs to GEC. The former compares ChatGPT to Grammarly and GECToR [93], a previous state-of-the-art GEC system, and the latter compares GPT-3.5[10] to two other GEC systems [94, 95]. Both approaches

---

[9]Note 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.

[10]Specifically, they use text-davinci-003.

find the GPT* models perform worse than existing systems when measured using automatic evaluation techniques on existing benchmark corpora (namely CoNLL-2014 [88], JFLEG [96], BEA-2019 [89]). The authors ascribe this to the model's tendency to *over-correct* learner text – by insertion of additional text or re-structuring of 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 found instances of *under*-correction 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, rather than told how to express it differently: that is, how to amend an error rather than avoid it [97]. The issue is not a new one [98] 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* [99] 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`[11]. 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'[12]) through careful feature extraction from parsers and GEC models, combined with prominent LLMs at the time such as T5 or GPT-Neo [100] (e.g. Babakov *et al.* [101] 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 [102].

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 [15] – 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,

---

[11]https://fcg.sharedtask.org/
[12]https://github.com/k-hanawa/fcg_genchal2022_baseline

inoffensive and preferably linked to teaching materials so that the learner may continue to benefit from EdTech applications for language.

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 is 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 have been trained in multilingual ways, the general trend with LLMs has exacerbated this pre-existing bias [103, 104, 105]. 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 [8, 9]. In addition, LLMs are known to exhibit certain biases [106] – both *representational* (language use around demographic groups) [107, 108] and *allocational* (how a system distributes resources or opportunities) [109, 107] – which need to be debiased or otherwise controlled [110, 111].

Suresh & Guttag identified various sources of harm in the 'machine learning life cycle' [109]: 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 [112, 113]. 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" [114, 115, 116]. 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.

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 [117]. 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 [118, 119, 120][13]. For example, Duolingo [121] sets out its approach to responsible AI under Validity & Reliability, Fairness, Privacy & Security, Accountability & Transparency

---

[13]https://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 [122] – 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 that involves 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, they are 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 [15], providing document-level assessment and feedback [123], handling code-switching or 'plurilingual' learning [124].

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional Transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[3] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018). URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[6] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., ChatGPT for good? On opportunities and challenges of large language models for education, Learning and Individual Differences 103 (2023) 102274. URL: https://doi.org/10.1016/j.lindif.2023.102274.

[7] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, U. Erlingsson, et al., Extracting training data from large language models., in: USENIX Security Symposium, volume 6, 2021.

[8] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019. URL: https://aclanthology.org/P19-1355. doi:10.18653/v1/P19-1355.

[9] J. Dodge, T. Prewitt, R. Tachet des Combes, E. Odmark, R. Schwartz, E. Strubell, A. S. Luccioni, N. A. Smith, N. DeCario, W. Buchanan, Measuring the carbon intensity of AI in cloud instances, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, 2022, p. 1877–1894. URL: https://doi.org/10.1145/3531146.3533234. doi:10.1145/3531146.3533234.

[10] C. Chou, L. Condron, J. C. Belland, A review of the research on Internet addiction, Educational Psychology Review 17 (2005) 363–388.

[11] V. R. Bhargava, M. Velasquez, Ethics of the attention economy: The problem of social media addiction, Business Ethics Quarterly 31 (2021) 321–359.

[12] F. Gioia, V. Rega, V. Boursier, Problematic internet use and emotional dysregulation among young people: A literature review., Clinical Neuropsychiatry: Journal of treatment Evaluation (2021).

[13] OpenAI, GPT-4 technical report arXiv:2303.08774 (2023). URL: https://arxiv.org/abs/2303.08774.

[14] W. Huang, K. F. Hew, L. K. Fryer, Chatbots for language learning—are they really useful? a systematic review of chatbot-supported language learning, Journal of Computer Assisted Learning 38 (2022) 237–257. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jcal.12610. doi:https://doi.org/10.1111/jcal.12610. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcal.12610.

[15] G. Tyen, M. Brenchley, A. Caines, P. Buttery, Towards an open-domain chatbot for language practice, in: Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), Seattle, Washington, 2022. URL: https://aclanthology.org/2022.bea-1.28. doi:10.18653/v1/2022.bea-1.28.

[16] Duolingo Team, Introducing Duolingo Max, a learning experience powered by gpt-4, 2023. URL: https://blog.duolingo.com/duolingo-max/.

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text Transformer, The Journal of Machine Learning Research 21 (2020).

[18] Google, PaLM 2 technical report, 2023. URL: https://ai.google/static/documents/palm2techreport.pdf.

[19] R. Thoppilan, D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, Q. Le, LaMDA: Language models for dialog applications arXiv:2201.08239 (2022). URL: https://arxiv.org/abs/2201.08239.

[20] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and efficient foundation language models arXiv:2302.13971 (2023). URL: https://arxiv.org/abs/2302.13971.

[21] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, OPT: Open Pre-trained Transformer language models arXiv:2205.01068 (2022). URL: https://arxiv.org/abs/2205.01068.

[22] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. van den Driessche, L. A. Hendricks, M. Rauh, P.-S. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen,

S. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J.-B. Lespiau, M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. de Masson d'Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. de Las Casas, A. Guy, C. Jones, J. Bradbury, M. Johnson, B. Hechtman, L. Weidinger, I. Gabriel, W. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, G. Irving, Scaling language models: Methods, analysis & insights from training Gopher arXiv:2112.11446 (2022). URL: https://arxiv.org/abs/2112.11446.

[23] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W. L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, P. Zhang, Y. Dong, J. Tang, GLM-130B: An open bilingual pre-trained model arXiv:2210.02414 (2022). URL: https://arxiv.org/abs/2210.02414.

[24] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, S. Weinbach, GPT-NeoX-20B: An open-source autoregressive language model, in: Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, Association for Computational Linguistics, virtual+Dublin, 2022, pp. 95–136. URL: https://aclanthology.org/2022.bigscience-1.9. doi:10.18653/v1/2022.bigscience-1.9.

[25] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, C. Leahy, The Pile: An 800GB dataset of diverse text for language modeling arXiv:2101.00027 (2020). URL: https://arxiv.org/abs/2101.00027.

[26] BigScience Workshop, :, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, D. Radev, E. G. Ponferrada, E. Levkovizh, E. Kim, E. B. Natan, F. D. Toni, G. Dupont, G. Kruszewski, G. Pistilli, H. Elsahar, H. Benyamina, H. Tran, I. Yu, I. Abdulmumin, I. Johnson, I. Gonzalez-Dios, J. de la Rosa, J. Chim, J. Dodge, J. Zhu, J. Chang, J. Frohberg, J. Tobing, J. Bhattacharjee, K. Almubarak, K. Chen, K. Lo, L. V. Werra, L. Weber, L. Phan, L. B. allal, L. Tanguy, M. Dey, M. R. Muñoz, M. Masoud, M. Grandury, M. Šaško, M. Huang, M. Coavoux, M. Singh, M. T.-J. Jiang, M. C. Vu, M. A. Jauhar, M. Ghaleb, N. Subramani, N. Kassner, N. Khamis, O. Nguyen, O. Espejel, O. de Gibert, P. Villegas, P. Henderson, P. Colombo, P. Amuok, Q. Lhoest, R. Harliman, R. Bommasani, R. L. López, R. Ribeiro, S. Osei, S. Pyysalo, S. Nagel, S. Bose, S. H. Muhammad, S. Sharma, S. Longpre, S. Nikpoor, S. Silberberg, S. Pai, S. Zink, T. T. Torrent, T. Schick, T. Thrush, V. Danchev, V. Nikoulina, V. Laippala, V. Lepercq, V. Prabhu, Z. Alyafeai, Z. Talat, A. Raja, B. Heinzerling, C. Si, D. E. Taşar, E. Salesky, S. J. Mielke, W. Y. Lee, A. Sharma, A. Santilli, A. Chaffin, A. Stiegler, D. Datta, E. Szczechla, G. Chhablani, H. Wang, H. Pandey, H. Strobelt, J. A. Fries, J. Rozen, L. Gao, L. Sutawika, M. S. Bari, M. S. Al-shaibani, M. Manica, N. Nayak, R. Teehan, S. Albanie, S. Shen, S. Ben-David, S. H.

Bach, T. Kim, T. Bers, T. Fevry, T. Neeraj, U. Thakker, V. Raunak, X. Tang, Z.-X. Yong, Z. Sun, S. Brody, Y. Uri, H. Tojarieh, A. Roberts, H. W. Chung, J. Tae, J. Phang, O. Press, C. Li, D. Narayanan, H. Bourfoune, J. Casper, J. Rasley, M. Ryabinin, M. Mishra, M. Zhang, M. Shoeybi, M. Peyrounette, N. Patry, N. Tazi, O. Sanseviero, P. von Platen, P. Cornette, P. F. Lavallée, R. Lacroix, S. Rajbhandari, S. Gandhi, S. Smith, S. Requena, S. Patil, T. Dettmers, A. Baruwa, A. Singh, A. Cheveleva, A.-L. Ligozat, A. Subramonian, A. Névéol, C. Lovering, D. Garrette, D. Tunuguntla, E. Reiter, E. Taktasheva, E. Voloshina, E. Bogdanov, G. I. Winata, H. Schoelkopf, J.-C. Kalo, J. Novikova, J. Z. Forde, J. Clive, J. Kasai, K. Kawamura, L. Hazan, M. Carpuat, M. Clinciu, N. Kim, N. Cheng, O. Serikov, O. Antverg, O. van der Wal, R. Zhang, R. Zhang, S. Gehrmann, S. Mirkin, S. Pais, T. Shavrina, T. Scialom, T. Yun, T. Limisiewicz, V. Rieser, V. Protasov, V. Mikhailov, Y. Pruksachatkun, Y. Belinkov, Z. Bamberger, Z. Kasner, A. Rueda, A. Pestana, A. Feizpour, A. Khan, A. Faranak, A. Santos, A. Hevia, A. Unldreaj, A. Aghagol, A. Abdollahi, A. Tammour, A. HajiHosseini, B. Behroozi, B. Ajibade, B. Saxena, C. M. Ferrandis, D. Contractor, D. Lansky, D. David, D. Kiela, D. A. Nguyen, E. Tan, E. Baylor, E. Ozoani, F. Mirza, F. Ononiwu, H. Rezanejad, H. Jones, I. Bhattacharya, I. Solaiman, I. Sedenko, I. Nejadgholi, J. Passmore, J. Seltzer, J. B. Sanz, L. Dutra, M. Samagaio, M. Elbadri, M. Mieskes, M. Gerchick, M. Akinlolu, M. McKenna, M. Qiu, M. Ghauri, M. Burynok, N. Abrar, N. Rajani, N. Elkott, N. Fahmy, O. Samuel, R. An, R. Kromann, R. Hao, S. Alizadeh, S. Shubber, S. Wang, S. Roy, S. Viguier, T. Le, T. Oyebade, T. Le, Y. Yang, Z. Nguyen, A. R. Kashyap, A. Palasciano, A. Callahan, A. Shukla, A. Miranda-Escalada, A. Singh, B. Beilharz, B. Wang, C. Brito, C. Zhou, C. Jain, C. Xu, C. Fourrier, D. L. Periñán, D. Molano, D. Yu, E. Manjavacas, F. Barth, F. Fuhrimann, G. Altay, G. Bayrak, G. Burns, H. U. Vrabec, I. Bello, I. Dash, J. Kang, J. Giorgi, J. Golde, J. D. Posada, K. R. Sivaraman, L. Bulchandani, L. Liu, L. Shinzato, M. H. de Bykhovetz, M. Takeuchi, M. Pàmies, M. A. Castillo, M. Nezhurina, M. Sänger, M. Samwald, M. Cullan, M. Weinberg, M. D. Wolf, M. Mihaljcic, M. Liu, M. Freidank, M. Kang, N. Seelam, N. Dahlberg, N. M. Broad, N. Muellner, P. Fung, P. Haller, R. Chandrasekhar, R. Eisenberg, R. Martin, R. Canalli, R. Su, R. Su, S. Cahyawijaya, S. Garda, S. S. Deshmukh, S. Mishra, S. Kiblawi, S. Ott, S. Sang-aroonsiri, S. Kumar, S. Schweter, S. Bharati, T. Laud, T. Gigant, T. Kainuma, W. Kusa, Y. Labrak, Y. S. Bajaj, Y. Venkatraman, Y. Xu, Y. Xu, Y. Xu, Z. Tan, Z. Xie, Z. Ye, M. Bras, Y. Belkada, T. Wolf, Bloom: A 176b-parameter open-access multilingual language model arXiv:2211.05100 (2023). URL: https://arxiv.org/abs/2211.05100.

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[28] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar,

S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, Y. Koreeda, Holistic Evaluation of Language Models arXiv:2211.09110 (2022). URL: https://arxiv.org/abs/2211.09110.

[29] Y. Park, G. T. LaFlair, Y. Attali, A. Runge, S. Goodwin, Duolingo English Test: Interactive reading (drr-22-02), 2022. URL: https://duolingo-papers.s3.amazonaws.com/other/mpr-whitepaper.pdf.

[30] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, Y. Wu, B. Neyshabur, G. Gur-Ari, V. Misra, Solving quantitative reasoning problems with language models arXiv:2206.14858 (2022). URL: https://arxiv.org/abs/2206.14858.

[31] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. URL: https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[32] L. Pan, W. Lei, T.-S. Chua, M.-Y. Kan, Recent advances in neural question generation, 2019. arXiv:1905.08949.

[33] V. Raina, M. Gales, Multiple-choice question generation: Towards an automated assessment framework, 2022. arXiv:2209.11830.

[34] M. Felice, S. Taslimipoor, P. Buttery, Constructing open cloze tests using generation and discrimination capabilities of Transformers, in: Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 2022. URL: https://aclanthology.org/2022.findings-acl.100. doi:10.18653/v1/2022.findings-acl.100.

[35] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, in: Proceedings of The International Conference on Learning Representations (ICLR), 2020.

[36] H.-L. Chung, Y.-H. Chan, Y.-C. Fan, A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies., in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4390–4400. URL: https://aclanthology.org/2020.findings-emnlp.393. doi:10.18653/v1/2020.findings-emnlp.393.

[37] S.-H. Chiang, S.-C. Wang, Y.-C. Fan, CDGP: Automatic cloze distractor generation based on pre-trained language model, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 5835–5840. URL: https://aclanthology.org/2022.findings-emnlp.429.

[38] P. Manakul, A. Liusie, M. J. F. Gales, MQAG: Multiple-choice question answering and generation for assessing information consistency in summarization, 2023. arXiv:2301.12307.

[39] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[40] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Barcelona, Spain, 2004. URL: https://aclanthology.org/W04-1013.

[41] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan, 2005. URL: https://aclanthology.org/W05-0909.

[42] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, 2020. arXiv:1904.09675.

[43] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, J. Pineau, Towards an automatic Turing Test: Learning to evaluate dialogue responses, 2018. arXiv:1708.07149.

[44] H. Shimanaka, T. Kajiwara, M. Komachi, RUSE: Regressor using sentence embeddings for automatic machine translation evaluation, in: Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Association for Computational Linguistics, Belgium, Brussels, 2018, pp. 751–758. URL: https://aclanthology.org/W18-6456. doi:10.18653/v1/W18-6456.

[45] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, RACE: Large-scale ReAding comprehension dataset from examinations, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 785–794. URL: https://aclanthology.org/D17-1082. doi:10.18653/v1/D17-1082.

[46] X. Kong, V. Gangal, E. Hovy, SCDE: Sentence cloze dataset with high quality distractors from examinations, 2020. arXiv:2004.12934.

[47] Q. Xie, G. Lai, Z. Dai, E. Hovy, Large-scale cloze test dataset created by teachers, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2344–2356. URL: https://aclanthology.org/D18-1257. doi:10.18653/v1/D18-1257.

[48] M. Felice, S. Taslimipoor, Ø. E. Andersen, P. Buttery, CEPOC: The Cambridge exams publishing open cloze dataset, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 4285–4290. URL: https://aclanthology.org/2022.lrec-1.456.

[49] A. Caines, H. Yannakoudakis, H. Allen, P. Pérez-Paredes, B. Byrne, P. Buttery, The Teacher-Student Chatroom Corpus version 2: more lessons, new annotation, automatic detection of sequence shifts, in: Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning, Louvain-la-Neuve, Belgium, 2022. URL: https://aclanthology.org/2022.nlp4call-1.3.

[50] T. Hosking, S. Riedel, Evaluating rewards for question generation models, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2278–2283. URL: https://aclanthology.org/N19-1237. doi:10.18653/v1/N19-1237.

[51] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback arXiv:2203.02155 (2022). URL: https://arxiv.org/abs/2203.02155.

[52] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano,

G. Irving, Fine-tuning language models from human preferences, 2020. `arXiv:1909.08593`.

[53] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, P. Christiano, Learning to summarize from human feedback, 2022. `arXiv:2009.01325`.

[54] L. Benedetto, P. Cremonesi, A. Caines, P. Buttery, A. Cappelli, A. Giussani, R. Turrin, A survey on recent approaches to question difficulty estimation from text, ACM Computing Surveys 55 (2023) 1–37.

[55] S. AlKhuzaey, F. Grasso, T. R. Payne, V. Tamma, A systematic review of data-driven approaches to item difficulty prediction, in: International Conference on Artificial Intelligence in Education, Springer, 2021, pp. 29–41.

[56] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, in: Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, 2020. URL: https://arxiv.org/abs/2210.02414.

[57] Y. Zhou, C. Tao, Multi-task bert for problem difficulty prediction, in: 2020 International Conference on Communications, Information System and Computer Engineering (CISCE), IEEE, 2020, pp. 213–216.

[58] L. Benedetto, G. Aradelli, P. Cremonesi, A. Cappelli, A. Giussani, R. Turrin, On the application of Transformers for estimating the difficulty of multiple-choice questions from text, in: Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, 2021, pp. 147–157.

[59] L. Benedetto, A quantitative study of nlp approaches to question difficulty estimation, arXiv preprint arXiv:2305.10236 (2023).

[60] E. Loginova, L. Benedetto, D. Benoit, P. Cremonesi, Towards the application of calibrated Transformers to the unsupervised estimation of question difficulty from text, in: RANLP 2021, INCOMA, 2021, pp. 846–855.

[61] L. Chen, K. Zechner, S.-Y. Yoon, K. Evanini, X. Wang, A. Loukina, J. Tao, L. Davis, C. M. Lee, M. Ma, R. Mundkowsky, C. Lu, C. W. Leong, B. Gyawali, Automated scoring of nonnative speech using the SpeechRaterSM v. 5.0 engine, ETS Research Report Series 2018 (2018) 1–31. doi:`https://doi.org/10.1002/ets2.12198`.

[62] K. Knill, M. Gales, K. Kyriakopoulos, A. Malinin, A. Ragni, Y. Wang, A. Caines, Impact of ASR performance on free speaking language assessment, in: Interspeech 2018, 2018, pp. 1641–1645.

[63] K. Zechner, K. Evanini, Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech, Routledge, 2019.

[64] H. Craighead, A. Caines, P. Buttery, H. Yannakoudakis, Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. URL: https://aclanthology.org/2020.acl-main.206. doi:`10.18653/v1/2020.acl-main.206`.

[65] J. Burstein, M. Chodorow, C. Leacock, Automated essay evaluation: The Criterion online writing service, AI Magazine 25 (2004) 27. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1774. doi:`10.1609/aimag.v25i3.1774`.

[66] Ø. E. Andersen, H. Yannakoudakis, F. Barker, T. Parish, Developing and testing a self-assessment and tutoring system, in: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, 2013. URL: https://aclanthology.org/

W13-1704.

[67] D. Alikaniotis, H. Yannakoudakis, M. Rei, Automatic text scoring using neural networks, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016. URL: https://aclanthology.org/P16-1068. doi:10.18653/v1/P16-1068.

[68] B. Riordan, A. Horbach, A. Cahill, T. Zesch, C. M. Lee, Investigating neural architectures for short answer scoring, in: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, 2017. URL: https://aclanthology.org/W17-5017. doi:10.18653/v1/W17-5017.

[69] Ø. E. Andersen, Z. Yuan, R. Watson, K. Y. F. Cheung, Benefits of alternative evaluation methods for automated essay scoring, in: Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021), Paris, France, 2021.

[70] H. Yannakoudakis, Ø. E. Andersen, A. Geranpayeh, T. Briscoe, D. Nicholls, Developing an automated writing placement system for ESL learners, Applied Measurement in Education 31 (2018) 251–267.

[71] R. Watson, E. Kochmar, Read & Improve: A novel reading tutoring system, in: Proceedings of Educational Data Mining (EDM), 2021.

[72] J. Jeon, S. Lee, Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT, Education and Information Technologies (2023) 1–20.

[73] A. Mizumoto, M. Eguchi, Exploring the potential of using an AI language model for automated essay scoring, Research Methods in Applied Linguistics 2 (2023) 100050.

[74] D. Blanchard, J. Tetreault, D. Higgins, A. Cahill, M. Chodorow, TOEFL11: A corpus of non-native English, ETS Research Report Series 2013 (2013) i–15.

[75] P. Lagakis, S. Demetriadis, Automated essay scoring: A review of the field, in: 2021 International Conference on Computer, Information and Telecommunication Systems (CITS), IEEE, 2021, pp. 1–6.

[76] European Commission, Regulatory framework proposal on artificial intelligence, 2022. URL: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[77] J. Wei, X. Wang, D. Schuurmans, M. Bosma, brian ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, in: A. H. Oh, A. Agarwal, D. Belgrave, K. Cho (Eds.), Advances in Neural Information Processing Systems, 2022. URL: https://openreview.net/forum?id=_VjQlMeSB_J.

[78] H. Yannakoudakis, T. Briscoe, B. Medlock, A new dataset and method for automatically grading ESOL texts, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, 2011. URL: https://aclanthology.org/P11-1019.

[79] A. Boyd, Using Wikipedia Edits in Low Resource Grammatical Error Correction, in: Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, Brussels, Belgium, 2018. URL: https://aclanthology.org/W18-6111. doi:10.18653/v1/W18-6111.

[80] J. Náplava, M. Straka, Grammatical Error Correction in Low-Resource Scenarios, in: Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), Hong Kong, China, 2019. URL: https://aclanthology.org/D19-5545. doi:10.18653/v1/D19-5545.

[81] O. Syvokon, O. Nahorna, P. Kuchmiichuk, N. Osidach, UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language, in: Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), Dubrovnik, Croatia, 2023. URL: https://aclanthology.org/2023.unlp-1.12.

[82] D. Dahlmeier, H. T. Ng, A beam-search decoder for grammatical error correction, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 568–578.

[83] S. Bell, H. Yannakoudakis, M. Rei, Context is key: Grammatical error detection with contextual word representations, in: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Florence, Italy, 2019. URL: https://aclanthology.org/W19-4410. doi:10.18653/v1/W19-4410.

[84] A. Caines, C. Bentz, K. Knill, M. Rei, P. Buttery, Grammatical error detection in transcriptions of spoken English, in: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 2020. URL: https://aclanthology.org/2020.coling-main.195. doi:10.18653/v1/2020.coling-main.195.

[85] Z. Yuan, S. Taslimipoor, C. Davis, C. Bryant, Multi-class grammatical error detection for correction: A tale of two systems, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 2021. URL: https://aclanthology.org/2021.emnlp-main.687. doi:10.18653/v1/2021.emnlp-main.687.

[86] M. Qorib, S.-H. Na, H. T. Ng, Frustratingly easy system combination for grammatical error correction, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, 2022. URL: https://aclanthology.org/2022.naacl-main.143. doi:10.18653/v1/2022.naacl-main.143.

[87] H. T. Ng, S. M. Wu, Y. Wu, C. Hadiwinoto, J. Tetreault, The CoNLL-2013 shared task on grammatical error correction, in: Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, Sofia, Bulgaria, 2013. URL: https://aclanthology.org/W13-3601.

[88] H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, C. Bryant, The CoNLL-2014 shared task on grammatical error correction, in: Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, Baltimore, Maryland, 2014. URL: https://aclanthology.org/W14-1701. doi:10.3115/v1/W14-1701.

[89] C. Bryant, M. Felice, Ø. E. Andersen, T. Briscoe, The BEA-2019 shared task on grammatical error correction, in: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Florence, Italy, 2019. URL: https://aclanthology.org/W19-4406. doi:10.18653/v1/W19-4406.

[90] C. Bryant, Z. Yuan, M. R. Qorib, H. Cao, H. T. Ng, T. Briscoe, Grammatical Error Correction: A Survey of the State of the Art, Computational Linguistics (2023) 1–59. URL: https://doi.org/10.1162/coli_a_00478. doi:10.1162/coli_a_00478.

[91] H. Wu, W. Wang, Y. Wan, W. Jiao, M. Lyu, ChatGPT or Grammarly? Evaluating ChatGPT on grammatical error correction benchmark arXiv:2303.13648 (2023). URL: https://arxiv.org/abs/2303.13648.

[92] S. Coyne, K. Sakaguchi, An analysis of GPT-3's performance in grammatical error

correction arXiv:2303.14342 (2023). URL: https://arxiv.org/abs/2303.14342.

[93] K. Omelianchuk, V. Atrasevych, A. Chernodub, O. Skurzhanskyi, Gector–grammatical error correction: tag, not rewrite, arXiv preprint arXiv:2005.12592 (2020).

[94] M. Yasunaga, J. Leskovec, P. Liang, Lm-critic: language models for unsupervised grammatical error correction, arXiv preprint arXiv:2109.06822 (2021).

[95] Z. Liu, X. Yi, M. Sun, L. Yang, T.-S. Chua, Neural quality estimation with multiple hypotheses for grammatical error correction, arXiv preprint arXiv:2105.04443 (2021).

[96] C. Napoles, K. Sakaguchi, J. Tetreault, JFLEG: A fluency corpus and benchmark for grammatical error correction, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, 2017. URL: https://aclanthology.org/E17-2037.

[97] D. Nicholls, The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT, in: Proceedings of the Corpus Linguistics 2003 conference; UCREL technical paper number 16, 2003. URL: https://ucrel.lancs.ac.uk/publications/cl2003/papers/nicholls.pdf.

[98] K. Sakaguchi, C. Napoles, M. Post, J. Tetreault, Reassessing the goals of grammatical error correction: Fluency instead of grammaticality, Transactions of the Association for Computational Linguistics 4 (2016). URL: https://aclanthology.org/Q16-1013. doi:10.1162/tacl_a_00091.

[99] R. Nagata, M. Hagiwara, K. Hanawa, M. Mita, A. Chernodub, O. Nahorna, Shared task on feedback comment generation for language learners, in: Proceedings of the 14th International Conference on Natural Language Generation, Aberdeen, Scotland, UK, 2021. URL: https://aclanthology.org/2021.inlg-1.35.

[100] S. Black, L. Gao, P. Wang, C. Leahy, S. Biderman, GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow, 2021. URL: https://doi.org/10.5281/zenodo.5297715. doi:10.5281/zenodo.5297715.

[101] N. Babakov, M. Lysyuk, A. Shvets, L. Kazakova, A. Panchenko, Error syntax aware augmentation of feedback comment generation dataset, in: Proceedings of the 16th International Natural Language Generation Conference, 2023. URL: https://arxiv.org/abs/2212.14293.

[102] S. Coyne, Template-guided grammatical error feedback comment generation, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, Dubrovnik, Croatia, 2023. URL: https://aclanthology.org/2023.eacl-srw.10.

[103] S. Gururangan, D. Card, S. Dreier, E. Gade, L. Wang, Z. Wang, L. Zettlemoyer, N. A. Smith, Whose language counts as high quality? measuring language ideologies in text data selection, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 2022. URL: https://aclanthology.org/2022.emnlp-main.165.

[104] A. Søgaard, Should we ban English NLP for a year?, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 2022. URL: https://aclanthology.org/2022.emnlp-main.351.

[105] K. Ramesh, S. Sitaram, M. Choudhury, Fairness in language models beyond English: Gaps and challenges, in: Findings of the Association for Computational Linguistics: EACL

2023, Dubrovnik, Croatia, 2023. URL: https://aclanthology.org/2023.findings-eacl.157.

[106] S. Barocas, K. Crawford, A. Shapiro, H. Wallach, The problem with bias: from allocative to representational harms in machine learning, in: Proceedings of the 9th Conference of the Special Interest Group for Computing, Information and Society (SIGCIS), 2017. URL: http://meetings.sigcis.org/uploads/6/3/6/8/6368912/program.pdf.

[107] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of "bias" in NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. URL: https://aclanthology.org/2020.acl-main.485. doi:10.18653/v1/2020.acl-main.485.

[108] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[109] H. Suresh, J. Guttag, A framework for understanding sources of harm throughout the machine learning life cycle, in: Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO), 2021. URL: https://doi.org/10.1145/3465416.3483305.

[110] M. Kaneko, D. Bollegala, Debiasing pre-trained contextualised embeddings, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021. URL: https://aclanthology.org/2021.eacl-main.107. doi:10.18653/v1/2021.eacl-main.107.

[111] J. Lalor, Y. Yang, K. Smith, N. Forsgren, A. Abbasi, Benchmarking intersectional biases in NLP, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, 2022. URL: https://aclanthology.org/2022.naacl-main.263. doi:10.18653/v1/2022.naacl-main.263.

[112] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, D. Roth, A comparative study of fairness-enhancing interventions in machine learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, 2019. URL: https://doi.org/10.1145/3287560.3287589. doi:10.1145/3287560.3287589.

[113] J. Finocchiaro, R. Maio, F. Monachou, G. K. Patro, M. Raghavan, A.-A. Stoica, S. Tsirtsis, Bridging machine learning and mechanism design towards algorithmic fairness, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, 2021. URL: https://doi.org/10.1145/3442188.3445912. doi:10.1145/3442188.3445912.

[114] N. Dehouche, Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3), Ethics in Science and Environmental Politics 21 (2021) 17–23.

[115] D. R. E. Cotton, P. A. Cotton, J. R. Shipway, Chatting and cheating: Ensuring academic integrity in the era of ChatGPT, Innovations in Education and Teaching International (2023) 1–12. URL: https://doi.org/10.1080/14703297.2023.2190148. doi:10.1080/14703297.2023.2190148.

[116] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, T. Goldstein, A watermark for large language models arXiv:2301.10226 (2023). URL: https://arxiv.org/abs/2301.10226.

[117] E. Santoro, B. Monin, The AI Effect: People rate distinctively human attributes as more essential to being human after learning about artificial intelligence advances, Journal of

Experimental Social Psychology 107 (2023) 104464. doi:https://doi.org/10.1016/j.jesp.2023.104464.

[118] V. Prabhakaran, M. Mitchell, T. Gebru, I. Gabriel, A human rights-based approach to responsible AI arXiv:2210.02667 (2022). URL: https://arxiv.org/abs/2210.02667, presented as a (non-archival) poster at the 2022 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization or (EAAMO '22).

[119] D. Thakkar, A. Ismail, P. Kumar, A. Hanna, N. Sambasivan, N. Kumar, When is machine learning data good?: Valuing in public health datafication, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, 2022. URL: https://doi.org/10.1145/3491102.3501868. doi:10.1145/3491102.3501868.

[120] M. Khan, A. Hanna, The subjects and stages of AI dataset development: A framework for dataset accountability, Ohio State Tech Law Journal (2023). URL: https://ssrn.com/abstract=4217148. doi:10.2139/ssrn.4217148.

[121] J. Burstein, K. Yancey, K. Bicknell, C. Gottlieb, M. Zheng, A. von Davier, Responsible AI standards, 2023. URL: https://duolingo-papers.s3.amazonaws.com/other/DET+Responsible+AI+033123.pdf.

[122] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, A. Xiang, A. Parrish, A. Nie, A. Hussain, A. Askell, A. Dsouza, A. Slone, A. Rahane, A. S. Iyer, A. Andreassen, A. Madotto, A. Santilli, A. Stuhlmüller, A. Dai, A. La, A. Lampinen, A. Zou, A. Jiang, A. Chen, A. Vuong, A. Gupta, A. Gottardi, A. Norelli, A. Venkatesh, A. Gholamidavoodi, A. Tabassum, A. Menezes, A. Kirubarajan, A. Mullokandov, A. Sabharwal, A. Herrick, A. Efrat, A. Erdem, A. Karakaş, B. R. Roberts, B. S. Loe, B. Zoph, B. Bojanowski, B. Özyurt, B. Hedayatnia, B. Neyshabur, B. Inden, B. Stein, B. Ekmekci, B. Y. Lin, B. Howald, C. Diao, C. Dour, C. Stinson, C. Argueta, C. F. Ramírez, C. Singh, C. Rathkopf, C. Meng, C. Baral, C. Wu, C. Callison-Burch, C. Waites, C. Voigt, C. D. Manning, C. Potts, C. Ramirez, C. E. Rivera, C. Siro, C. Raffel, C. Ashcraft, C. Garbacea, D. Sileo, D. Garrette, D. Hendrycks, D. Kilman, D. Roth, D. Freeman, D. Khashabi, D. Levy, D. M. González, D. Perszyk, D. Hernandez, D. Chen, D. Ippolito, D. Gilboa, D. Dohan, D. Drakard, D. Jurgens, D. Datta, D. Ganguli, D. Emelin, D. Kleyko, D. Yuret, D. Chen, D. Tam, D. Hupkes, D. Misra, D. Buzan, D. C. Mollo, D. Yang, D.-H. Lee, E. Shutova, E. D. Cubuk, E. Segal, E. Hagerman, E. Barnes, E. Donoway, E. Pavlick, E. Rodola, E. Lam, E. Chu, E. Tang, E. Erdem, E. Chang, E. A. Chi, E. Dyer, E. Jerzak, E. Kim, E. E. Manyasi, E. Zheltonozhskii, F. Xia, F. Siar, F. Martínez-Plumed, F. Happé, F. Chollet, F. Rong, G. Mishra, G. I. Winata, G. de Melo, G. Kruszewski, G. Parascandolo, G. Mariani, G. Wang, G. Jaimovitch-López, G. Betz, G. Gur-Ari, H. Galijasevic, H. Kim, H. Rashkin, H. Hajishirzi, H. Mehta, H. Bogar, H. Shevlin, H. Schütze, H. Yakura, H. Zhang, H. M. Wong, I. Ng, I. Noble, J. Jumelet, J. Geissinger, J. Kernion, J. Hilton, J. Lee, J. F. Fisac, J. B. Simon, J. Koppel, J. Zheng, J. Zou, J. Kocoń, J. Thompson, J. Kaplan, J. Radom, J. Sohl-Dickstein, J. Phang, J. Wei, J. Yosinski, J. Novikova, J. Bosscher, J. Marsh, J. Kim, J. Taal, J. Engel, J. Alabi, J. Xu, J. Song, J. Tang, J. Waweru, J. Burden, J. Miller, J. U. Balis, J. Berant, J. Frohberg, J. Rozen, J. Hernandez-Orallo, J. Boudeman, J. Jones, J. B. Tenenbaum, J. S. Rule, J. Chua, K. Kanclerz, K. Livescu, K. Krauth, K. Gopalakrishnan, K. Ignatyeva, K. Markert, K. D. Dhole, K. Gimpel, K. Omondi, K. Mathewson, K. Chiafullo,

K. Shkaruta, K. Shridhar, K. McDonell, K. Richardson, L. Reynolds, L. Gao, L. Zhang, L. Dugan, L. Qin, L. Contreras-Ochando, L.-P. Morency, L. Moschella, L. Lam, L. Noble, L. Schmidt, L. He, L. O. Colón, L. Metz, L. K. Şenel, M. Bosma, M. Sap, M. ter Hoeve, M. Farooqi, M. Faruqui, M. Mazeika, M. Baturan, M. Marelli, M. Maru, M. J. R. Quintana, M. Tolkiehn, M. Giulianelli, M. Lewis, M. Potthast, M. L. Leavitt, M. Hagen, M. Schubert, M. O. Baitemirova, M. Arnaud, M. McElrath, M. A. Yee, M. Cohen, M. Gu, M. Ivanitskiy, M. Starritt, M. Strube, M. Swędrowski, M. Bevilacqua, M. Yasunaga, M. Kale, M. Cain, M. Xu, M. Suzgun, M. Tiwari, M. Bansal, M. Aminnaseri, M. Geva, M. Gheini, M. V. T, N. Peng, N. Chi, N. Lee, N. G.-A. Krakover, N. Cameron, N. Roberts, N. Doiron, N. Nangia, N. Deckers, N. Muennighoff, N. S. Keskar, N. S. Iyer, N. Constant, N. Fiedel, N. Wen, O. Zhang, O. Agha, O. Elbaghdadi, O. Levy, O. Evans, P. A. M. Casares, P. Doshi, P. Fung, P. P. Liang, P. Vicol, P. Alipoormolabashi, P. Liao, P. Liang, P. Chang, P. Eckersley, P. M. Htut, P. Hwang, P. Miłkowski, P. Patil, P. Pezeshkpour, P. Oli, Q. Mei, Q. Lyu, Q. Chen, R. Banjade, R. E. Rudolph, R. Gabriel, R. Habacker, R. R. Delgado, R. Millière, R. Garg, R. Barnes, R. A. Saurous, R. Arakawa, R. Raymaekers, R. Frank, R. Sikand, R. Novak, R. Sitelew, R. LeBras, R. Liu, R. Jacobs, R. Zhang, R. Salakhutdinov, R. Chi, R. Lee, R. Stovall, R. Teehan, R. Yang, S. Singh, S. M. Mohammad, S. Anand, S. Dillavou, S. Shleifer, S. Wiseman, S. Gruetter, S. R. Bowman, S. S. Schoenholz, S. Han, S. Kwatra, S. A. Rous, S. Ghazarian, S. Ghosh, S. Casey, S. Bischoff, S. Gehrmann, S. Schuster, S. Sadeghi, S. Hamdan, S. Zhou, S. Srivastava, S. Shi, S. Singh, S. Asaadi, S. S. Gu, S. Pachchigar, S. Toshniwal, S. Upadhyay, Shyamolima, Debnath, S. Shakeri, S. Thormeyer, S. Melzi, S. Reddy, S. P. Makini, S.-H. Lee, S. Torene, S. Hatwar, S. Dehaene, S. Divic, S. Ermon, S. Biderman, S. Lin, S. Prasad, S. T. Piantadosi, S. M. Shieber, S. Misherghi, S. Kiritchenko, S. Mishra, T. Linzen, T. Schuster, T. Li, T. Yu, T. Ali, T. Hashimoto, T.-L. Wu, T. Desbordes, T. Rothschild, T. Phan, T. Wang, T. Nkinyili, T. Schick, T. Kornev, T. Telleen-Lawton, T. Tunduny, T. Gerstenberg, T. Chang, T. Neeraj, T. Khot, T. Shultz, U. Shaham, V. Misra, V. Demberg, V. Nyamai, V. Raunak, V. Ramasesh, V. U. Prabhu, V. Padmakumar, V. Srikumar, W. Fedus, W. Saunders, W. Zhang, W. Vossen, X. Ren, X. Tong, X. Zhao, X. Wu, X. Shen, Y. Yaghoobzadeh, Y. Lakretz, Y. Song, Y. Bahri, Y. Choi, Y. Yang, Y. Hao, Y. Chen, Y. Belinkov, Y. Hou, Y. Hou, Y. Bai, Z. Seid, Z. Zhao, Z. Wang, Z. J. Wang, Z. Wang, Z. Wu, Beyond the imitation game: Quantifying and extrapolating the capabilities of language models arXiv:2206.04615 (2022). URL: https://arxiv.org/abs/2206.04615.

[123] T. Wambsganss, A. Caines, P. Buttery, ALEN app: Argumentative writing support to foster English language learning, in: Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), Seattle, Washington, 2022. URL: https://aclanthology.org/2022.bea-1.18. doi:10.18653/v1/2022.bea-1.18.

[124] L. Nguyen, Z. Yuan, G. Seed, Building educational technologies for code-switching: Current practices, difficulties and future directions, Languages 7 (2022) 220.