Received: 3 June 2024 | Accepted: 26 June 2024

DOI: 10.1111/bjet.13505

REVIEW ARTICLE

British Journal of Educational Technology BERA

The life cycle of large language models in education: A framework for understanding sources of bias

Jinsook Lee[1] | **Yann Hicke**[2] | **Renzhe Yu**[3] | **Christopher Brooks**[4] | **René F. Kizilcec**[1]

[1]Department of Information Science, Cornell University, Ithaca, New York, USA

[2]Department of Computer Science, Cornell University, Ithaca, New York, USA

[3]Teachers College and Data Science Institute, Columbia University, New York, New York, USA

[4]School of Information, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Jinsook Lee, Department of Information Science, Cornell University, Ithaca, NY, USA.
Email: jl3369@cornell.edu

Funding information
Learning Engineering Virtual Institute (LEVI); Jacobs Foundation Research Fellowship

Abstract: Large language models (LLMs) are increasingly adopted in educational contexts to provide personalized support to students and teachers. The unprecedented capacity of LLM-based applications to understand and generate natural language can potentially improve instructional effectiveness and learning outcomes, but the integration of LLMs in education technology has renewed concerns over algorithmic bias, which may exacerbate educational inequalities. Building on prior work that mapped the traditional machine learning life cycle, we provide a framework of the LLM life cycle from the initial development of LLMs to customizing pre-trained models for various applications in educational settings. We explain each step in the LLM life cycle and identify potential sources of bias that may arise in the context of education. We discuss why current measures of bias from traditional machine learning fail to transfer to LLM-generated text (eg, tutoring conversations) because text encodings are high-dimensional, there can be multiple correct responses, and tailoring responses may be pedagogically desirable rather than unfair. The proposed framework clarifies the complex nature of bias in LLM applications and provides practical guidance for their evaluation to promote educational equity.

KEYWORDS
bias and fairness, education, large language model (LLM)

Practitioner notes

What is already known about this topic

- The life cycle of traditional machine learning (ML) applications which focus on predicting labels is well understood.
- Biases are known to enter in traditional ML applications at various points in the life cycle, and methods to measure and mitigate these biases have been developed and tested.
- Large language models (LLMs) and other forms of generative artificial intelligence (GenAI) are increasingly adopted in education technologies (EdTech), but current evaluation approaches are not specific to the domain of education.

What this paper adds

- A holistic perspective of the LLM life cycle with domain-specific examples in education to highlight opportunities and challenges for incorporating natural language understanding (NLU) and natural language generation (NLG) into EdTech.
- Potential sources of bias are identified in each step of the LLM life cycle and discussed in the context of education.
- A framework for understanding where to expect potential harms of LLMs for students, teachers, and other users of GenAI technology in education, which can guide approaches to bias measurement and mitigation.

Implications for practice and/or policy

- Education practitioners and policymakers should be aware that biases can originate from a multitude of steps in the LLM life cycle, and the life cycle perspective offers them a heuristic for asking technology developers to explain each step to assess the risk of bias.
- Measuring the biases of systems that use LLMs in education is more complex than with traditional ML, in large part because the evaluation of natural language generation is highly context-dependent (eg, what counts as good feedback on an assignment varies).
- EdTech developers can play an important role in collecting and curating datasets for the evaluation and benchmarking of LLM applications moving forward.

INTRODUCTION

In late 2022, large language models (LLMs) and generative artificial intelligence (AI) captured widespread attention when OpenAI released a public beta version of its LLM-based chatbot ChatGPT. It offered a compelling demonstration of the state of the art in generative AI chatbots by engaging in text-based conversations that exhibit forms of intelligence and a human-like tone. The technology was put to the test, quite literally, and scored extremely highly on a large variety of standardized tests, in addition to fooling a panel of judges in a version of the Turing test, which led scientists to question the validity of the famous benchmark for machine intelligence (Biever, 2023). Realizing the immense impact that LLMs can have in education, OpenAI partnered with Khan Academy ahead of the public release of GPT-4 to help the EdTech provider integrate a version of GPT-4 into its learning platform as an "AI-powered guide, tutor for learners, and assistant for teachers" called Khanmigo (Khan Academy, n.d.). Similar AI-powered learning assistants quickly appeared in other major

EdTech platforms, such as Coach on the Coursera platform (Coursera, 2023) and XPert on the EdX platform (edX Press, n.d.). These chatbots are perhaps the closest anyone has come to a scalable and domain-agnostic solution to Bloom's Two-Sigma Problem on how to provide large numbers of learners with support that is as effective as personal tutoring using a mastery-learning approach (Bloom, 1984). EdTech providers are developing new features using LLMs to enhance their products, including AI tutors that answer student questions in real-time, provide instant, personalized feedback on written assignments, or help teachers create new assignments and grade them faster with detailed feedback. There are numerous potential applications of this new technology in education (Yan et al., 2024), which raises questions about the long-term impacts of AI in education, and more immediate questions about issues that can arise when AI-based technology, built on data sourced from the World Wide Web, is deployed in classrooms (Denny et al., 2024; Yan et al., 2024).

In this article, we focus on the potential biases that LLMs may exhibit in the context of education. Algorithmic biases tend to negatively impact members of disadvantaged groups and perpetuate inequities at a larger scale. Most LLMs, including GPT models (Brown et al., 2020; Bubeck et al., 2023; OpenAI, 2023; Radford et al., 2019), Palm 2 (Anil et al., 2023), BLOOM (Workshop et al., 2023), LLaMA (Touvron et al., 2023), Flan-T5 (Chung et al., 2024), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), are trained on extremely large web corpora, which can cause them to learn social biases even when active steps are taken to mitigate them. This can be difficult to examine directly because many LLMs, including those developed by OpenAI, are not released as open-source models and provide limited information on how models were trained and evaluated. A growing number of open-weight models have been released, including Mistral (Jiang et al., 2023), Falcon (Almazrouei et al., 2023), Gemma (Gemma Team, Mesnard, et al., 2024), Gemini (Gemini Team Google, Anil, et al., 2024) and QWen (Bai et al., 2023). These models allow the community to study their fairness properties more closely, providing valuable insights into their performance and biases. However, many of these models are not fully open-source, as they do not provide comprehensive details such as the model architecture code, training methodology, hyperparameters, original training datasets, documentation and other relevant information. Additionally, biases can also arise based on how models are integrated into an application, which has sparked efforts to promote responsible AI using application-specific licensing (eg, the BigScience RAIL Licence[1]).

The rapid adoption of LLM-based technology in educational institutions presses the need to systematically evaluate LLMs for bias to avoid unintended consequences, such as amplifying current educational inequities in opportunity and achievement. Although there is an established area of research on AI bias and fairness, including a domain-specific literature for education (Baker & Hawn, 2022; Kizilcec & Lee, 2022), there is limited guidance on what potential biases can arise in the process of LLM development, how to evaluate and mitigate bias in LLM-based applications, specifically in the context of education. Applications of LLM-based generative AI raise particular challenges for evaluating bias due to the complexity of its natural language output and establishing a ground truth that is appropriate for the context of use. This article aims to improve our understanding of bias resulting from LLMs in educational applications. To define the context of these applications, we first present a set of studies that use LLM technology to support a variety of tasks in educational settings. Then, building on an established framework of the life cycle of (traditional) machine learning, we propose a new framework of the LLM life cycle that traces each step from the initial development to the final touches of customization for LLM-based applications. For each step in the LLM life cycle, we highlight potential biases that can arise in educational contexts and potential measures of those biases. We discuss the implications of the LLM life cycle for researchers interested in evaluating and mitigating bias, practitioners interested in understanding where biases might arise from and policymakers looking to better understand the

ethical issues related to LLM use in education, including the opportunity cost of not using LLMs. Evaluating this opportunity cost is crucial; while addressing and mitigating bias in LLMs is important, abandoning them without considering their benefits could hinder educational progress by depriving students of valuable feedback and support. In this article, we highlight opportunities and practical challenges of using LLMs in education and important areas for future research on LLM bias and fairness in education.

LLM APPLICATIONS IN EDUCATION

There are a variety of ways that LLMs can be used in educational contexts, many of which have been described by Yan et al. (2024). We organize them into two broad types of use cases: *natural language generation* (NLG) and *natural language understanding* (NLU) tasks. NLG tasks include creating educational content, such as lesson plans, assessments, and in-class materials like worksheets (Kasneci et al., 2023; Leiker et al., 2023; Wollny et al., 2021). NLU tasks involve analysing text for an educational purpose, such as making a prediction based on a student's essay submission about how well they understood the materials and scored on a given grading rubric. NLU tasks can also serve as an input into a larger model, such as an LLM used to detect confusion in a student's question, which can serve as an input into a predictive model for student underperformance and drop-out. An NLU task can also serve as the first step of an NLG task: an AI-based grading system, for example, may first analyse and score a student's essay and then generate written feedback based on that analysis (Zheng et al., 2022). Other examples of combined NLU-NLG tasks include tutoring chatbots like Khanmigo (Khan Academy, n.d.) and Rori (Henkel et al., 2024), which provide customized guidance to students across subjects including mathematics and the language arts, systems that provide personalized hints for compiler errors in a programming course (Pankiewicz & Baker, 2024), and tools designed to provide feedback or training to educators and tutors (Lin et al., 2023). In the context of this review article, we focus on cases where LLMs are used to enhance teaching and learning, and we therefore do not consider use cases like LLM-based essay-writing services.

In the responsible AI literature, algorithmic biases have been organized into two broad categories: representational biases and allocative biases (Suresh & Guttag, 2021). The potential biases associated with NLG tasks are mostly representational biases because NLG tasks can create text containing stereotypes or misrepresentations, exclusionary language or even toxic content (Weidinger et al., 2021).[2] The potential biases associated with NLU tasks are mostly allocative biases (ie, a bias in the allocation of resources) because individuals may receive differential access to resources or opportunities (Suresh & Guttag, 2021). For example, a grading system using an LLM for NLU could systematically assign lower scores to students of certain demographic groups, even though no identifying information was provided to the LLM. In fact, LLMs have been shown to display dialect prejudice when asked to make decisions about speakers of African American English (AAE) as compared to speakers of Standard American English (SAE) (Hofmann et al., 2024). Educational applications that rely on both NLU and NLG are susceptible to both types of biases. For example, an intelligent tutoring system might generate assessments that inadvertently reinforce stereotypes (a representational bias) and also disproportionately show those assessments to students with certain backgrounds (an allocative bias).

The classification of different tasks (NLU and NLG) and types of biases (allocative and representational) begin to organize the complexity associated with bias from LLMs in education technology. However, it does not explain where biases originate in the multi-step process from developing to customizing to ultimately deploying an LLM for an educational purpose. We therefore developed a framework for understanding this multi-step process to

help identify where biases might emerge, for what reasons and how to potentially measure them.

THE LLM LIFE CYCLE FROM DEVELOPMENT TO DEPLOYMENT

We build on the machine learning life cycle framework proposed by Suresh and Guttag (2021). It pinpoints where bias can be introduced in the process of creating and deploying a system using traditional machine learning. We have modified their original framework for the specific context of LLM-based applications, which is substantially more complex, to examine where bias may be introduced (Figures 1 and 2). Due to its complexity, we divide the life cycle into two phases: the initial development phase of the base LLM, and the customization phase which relies on a base LLM. We describe potential biases in each step of the life cycle with examples from education contexts.

Phase 1: Training a base LLM

Scraping and sampling

Large language models (LLMs) are trained using extensive text corpora, such as WebText or Common Crawl (Radford et al., 2019), which are scraped from pages on the World Wide Web. Online text data can reflect both current and past discrimination. Biases can arise from prejudices contained in these data, including biases inherent in the text (ie, the content of the text) or biases arising from the selection process (ie, which texts are included and which are excluded). *Historical bias* frequently arises when data

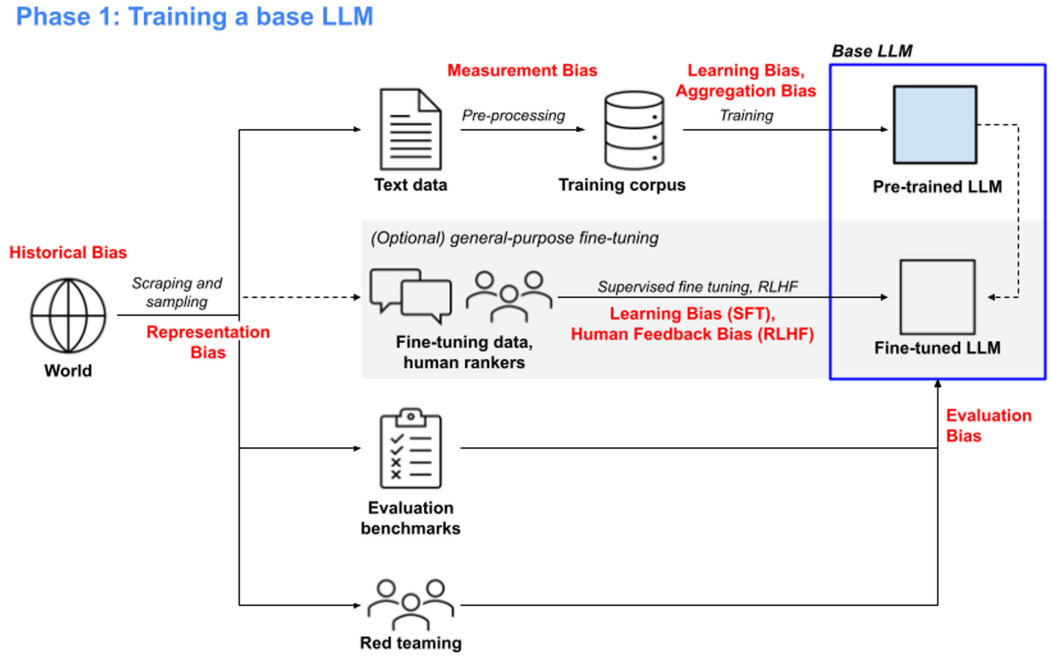

FIGURE 1 The initial development phase of the LLM life cycle with potential sources of biases, after Suresh and Guttag (2021).

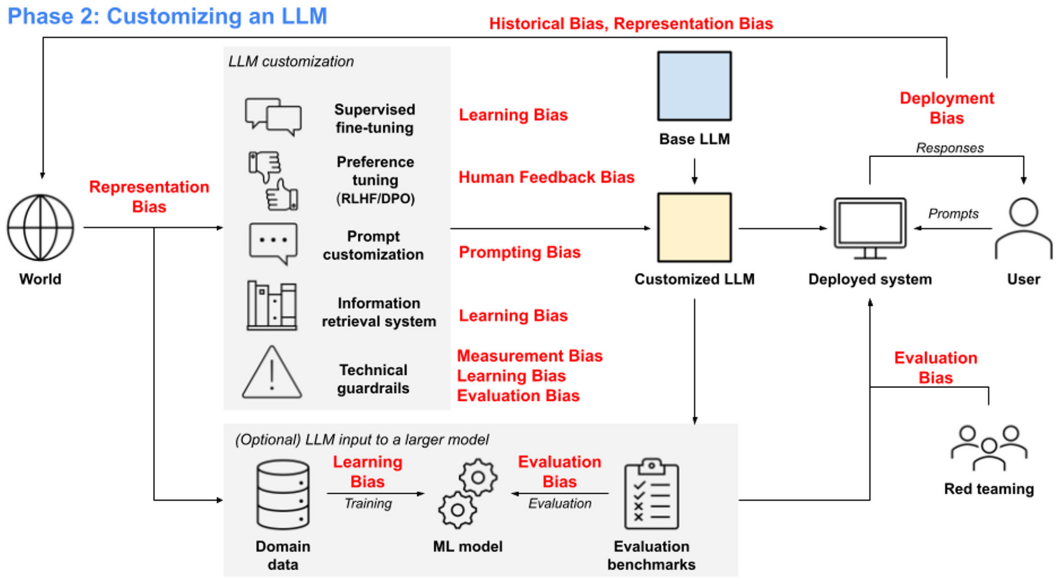

FIGURE 2 The customization phase of the LLM life cycle with potential sources of biases, after Suresh and Guttag (2021).

are collected over a long period and unintentionally reveals historical discrimination for certain groups. For instance, when collecting data related to STEM fields, there tends to be an imbalanced gender representation because there has historically been less representation of women in these areas. Additionally, due to the vast amount of data from various sources, genres and periods, the content may include discriminatory elements, such as documents involved in discrimination, which can pose harm to certain groups (Barocas & Selbst, 2016).

Considering the historical biases that have accumulated globally, *representation bias* can emerge in the form of an imbalance in the sampled data along dimensions including language, sample periods, available sources, and authorship. Ultimately, there is no way to avoid these difficult choices during the sampling process to narrow down the vast volume and diversity of text on the Internet. Representation bias can arise due to source availability and related policy restrictions, resulting in a predominant collection of English-focused datasets, while datasets for other languages could be relatively underrepresented. Consequently, content in other languages might not be fully represented in the actual world. Additionally, the choice of when to start scraping and sampling can cause representation bias because data gathered a long time ago might not reflect present-day conditions. The consequences of representation bias, including geographical (Ocumpaugh et al., 2014) and temporal (Levin et al., 2022) bias in the training data, have been examined in the context of education technology. Yet representation bias can occur not only during data sampling but also when recruiting people for data labelling or 'red teaming' (the practice of recruiting an external team to discover risks by taking an adversarial approach, for example, showing biases by trying to elicit them from the system). The background characteristics of individuals recruited for these efforts can present a further source of representation bias.

The unregulated nature of World Wide Web content can further contribute to representation bias. Specifically, *harmful content* that is explicitly or implicitly stereotyping, misrepresenting, and using toxic or exclusionary language can affect the representation of members of certain groups in the training corpus. A number of open training datasets, such as *LAION-400M* (T. L. Team, 2024), have been found to contain disturbing and explicit

content, including images-text pairs related to rape, pornography, harmful stereotypes, as well as racist and derogatory remarks about some ethnic backgrounds (Birhane et al., 2021, 2023). This evidence suggests that larger-scale versions of these datasets could exacerbate representational bias.

Once a text corpus for developing the LLM has been sampled, the next step is to *pre-process* the data. To improve data quality, duplicate texts are removed, noisy data are removed (eg, very short pieces of text), personally identifiable information is removed or masked, texts related to popular benchmarks are removed to ensure a fair evaluation (a process known as decontamination), and texts containing toxic or overtly biased language are removed (Weights & Biases, 2023). The process of filtering out toxic and biased content relies on dictionaries (LDNOOBW, 2023) or detection tools (spamscanner, 2023). However, these may not capture all instances of objectionable speech. We can apply a framework to help parse harmful content, for instance, by categorizing it along the type of harm (eg, misinformation, hate speech, stereotypes), whether harmful content is sought out for the specific application (eg, to learn how to identify it better going forward) or not, and who is affected by the harmful content (eg, individuals represented in the dataset, demographic groups) (Kirk et al., 2022).

In developing tools or frameworks to process raw text data, we may inadvertently encounter *measurement bias*, defined here as a systematic error in measuring specific abstract concepts (eg, toxicity, bias, private information). A feature typically represents a specific measurement that stands in for a broader and often intangible concept. For example, it can be challenging to measure the concept of "toxic" when there are only subtle and implicit discriminatory words in the text. If certain slang terms are commonly used by a particular community, it can be difficult to determine whether the words are toxic or not. Measurement bias can also arise from people tasked with identifying instances of the construct. The opinions of individuals who label toxic and biased content are shaped by their viewpoint and background, which can reinforce their perspectives (and exclude others) through the process of data curation and filtering (Weights & Biases, 2023). Overall, this inherent ambiguity in quantifying abstract constructs can introduce measurement bias when operationalizing these constructs during data pre-processing, and ultimately lead to harm (Jacobs & Wallach, 2021).

Pre-training (training corpus → pre-trained LLM)

Once the training corpus is pre-processed, the next step is *tokenization*. The text data is broken down into pieces that can be words, parts of words, or byte pairs. This process transforms the text corpus into a format that models can process. These are used to create word and contextual embeddings to represent features that allow machine learning models to easily correlate input data with output data. These embeddings are designed to capture the semantic and syntactic properties of words within a high-dimensional space, thereby enhancing the model's capability in NLU and NLG tasks. This sets up the architecture for pre-training the model, which involves a sequence of transformer blocks with multi-head self-attention mechanisms and fully connected layers of neural networks (Radford & Narasimhan, 2018).

Once this architecture is set up, the model is pre-trained to predict the next token in sequence, and during the pre-training, the model's weights are optimized while its predictions are continuously compared to the actual outcomes, using the errors to update the weights in each step. The model thereby learns contextualized representations of words and phrases. Typically, the loss function used for pre-training LLMs is cross-entropy loss, which measures the difference between the predicted probability distribution and the

distribution of the actual next token (Mehrabi et al., 2021; Minaee et al., 2024). Pre-training techniques vary based on whether the focus is on NLU or NLG. For NLU, models like BERT utilize masking techniques where some words in a sentence are hidden, and the model is trained to predict these masked words. This approach helps the model grasp the context and meaning of sentences. Additionally, BERT employs a next-sentence prediction task where the model predicts whether a sentence logically follows a given sentence, further enhancing its understanding capabilities. On the other hand, NLG-focused models like GPT are pre-trained using a next-word prediction task, where the model learns to predict the next word in a sequence given the previous words. This sequential prediction task is important for generating coherent and contextually relevant text (Solaiman & Dennison, 2021).

However, *learning bias* (also known as algorithmic bias) can arise during this process driven by an objective function like minimizing cross-entropy loss if undesirable biases in the training data are inadvertently amplified. We define learning bias in LLMs as *amplifying undesirable inherent biases* when there is a goal to minimize a given loss function. The bias that is encoded in this step can be considered intrinsic to the model because it resides in the geometry of the embedding space (Goldfarb-Tarrant et al., 2021). There is a plethora of studies examining learning bias in word and contextual embedding spaces, including ones that study gender bias (Bolukbasi et al., 2016; Zhao et al., 2019), gender and ethnic stereotypes (Garg et al., 2018), gender neural words (Zhao et al., 2019), cultural biases (Durrheim et al., 2023; Swinger et al., 2019; Tao et al., 2024), and studies that trace training documents to identify the origin of such biases (Brunet et al., 2019). There is also extensive research on debiasing word embeddings, such as reducing gender bias (Bolukbasi et al., 2016; Gonen & Goldberg, 2019).

Intrinsic learning bias can be measured with either embedding-based metrics or probability-based metrics. Embedding-based metrics are computed distances in the vector space between words/sentences representing the domain of evaluation (eg, professions) and words/sentences representing the identities being evaluated for bias (eg, genders, racial groups). The Word Embedding Association Test (WEAT) Caliskan et al. (2017) is a commonly used embedding-based metric that quantifies biases in word embedding by examining how closely words related to certain concepts are associated with words related to social groups or attributes. Likewise, the Sentence Encoder Association Test (SEAT) quantifies bias in a set of sentences by encoding them into numerical embeddings using a sentence encoder model (May et al., 2019). Probability-based metrics are computed based on the likelihood of predictions. For example, the Discovery of Correlations (DisCo) method Webster et al. (2020) uses masked tokens in a template sentence completion task. The first part of the template sentence includes a word related to a specific social group (eg, gendered names or pronouns), and the second part has the language model predict the top three words that might complete the sentence. DisCo counts how often the model predicts different words for different social groups across all templates to obtain a probability-based measure of bias. While DisCo focuses on uncovering patterns within the model's predictions, the Log-Probability Bias Score (LPBS) (Kurita et al., 2019) measures intrinsic probability distributions of the model's outputs by directly measuring how likely the model is to produce certain biased outputs based on the log-probabilities.

Finally, *aggregation bias* can arise when a chosen model does not perform equally well across all subgroups, often because the data includes distinct subgroups that are treated uniformly instead of individually (Hutiri & Ding, 2022). There may not be a one-size-fits-all model that does not make any sacrifice on performance for certain groups. This bias is relevant for both NLU and NLG tasks, for example, in that a model works well for one language but is not the optimal choice for other languages.

(Optional) general-purpose fine-tuning

After pre-training the language model, LLM developers may use general-purpose fine-tuning, which is the process of taking a pre-trained model and further training it on a more specific dataset or task. This can be achieved using supervised fine-tuning (SFT), which is a method for refining a pre-trained model using labelled data. SFT adapts the pre-trained model's parameters to behave in a certain way based on a dataset that provides concrete examples of how it should behave (ie, a supervised target) (Radford et al., 2018). Since this step adds an additional dataset, *representation biases* can be introduced here too. And since SFT updates the pre-trained model parameters based on a chosen objective function, *learning biases* can arise as well.

Reinforcement learning from human feedback (RLHF), which is a type of SFT, is increasingly used to fine-tune the model's behaviour to better align with the goals, needs or preferences of a user group (OpenAI, 2023). Human raters are recruited to provide a large number of rankings of text outputs based on criteria such as harmlessness and helpfulness (Bai et al., 2022). The resulting dataset contains important signals for what output is more desirable for a particular user group, a domain or a task, but *human feedback bias* can be introduced in this step. Human feedback bias creates issues when these ratings mistakenly reinforce a model to behave in undesirable ways. The RLHF process requires high-quality feedback data, and undesirable outcomes can occur if the instructions provided during the labelling process are insufficient or unclear. For example, without proper guidance and training, human raters might generate preference data that leads the model to suggest harmful actions, such as criminal activity (OpenAI, 2023).

Beyond human feedback bias, *representation bias*, *measurement bias* and *learning bias* can also emerge during RLHF. Representation bias can arise if the sample characteristics of the human raters do not adequately represent the relevant population of the model's application context. Measurement bias can arise because concepts like harmlessness and helpfulness are abstract, and human raters might have varying standards in mind when making judgements. Learning bias can occur during the process of updating model parameters, depending on how the reward model is created and the objective function is chosen. The challenges and open problems associated with human feedback bias in RLHF include that human raters may pursue incorrect and harmful goals, including giving adversarial ratings that are hard to spot but that can lead to data poisoning (Casper et al., 2023).

Base LLM evaluation

Before the base LLM is ready, it needs to undergo an evaluation step. Many benchmark datasets have been created for the purpose of evaluating LLMs by testing different aspects of the model's capacities on NLU and NLG tasks. In addition to standard benchmark datasets, some developers may engage a group of critical external testers to find flaws or vulnerabilities in a model's performance and behaviour–a process known as red teaming. This adversarial approach helps uncover potential weaknesses that might not be evident through standard evaluation methods. By actively trying to break the model or cause it to produce incorrect or biased outputs, the red teaming provides valuable insights into the model's robustness and safety (Ganguli et al., 2022). A popular framework for evaluating LLMs is the Holistic Evaluation of Language Models (HELM) project (Liang et al., 2023), which includes a number of evaluations that focus on the interpretability and transparency of models, including bias metrics such as toxicity.

Evaluation bias can arise in this step because there are many choices for evaluating the model, which can lead to substantially different conclusions. First, since benchmark

datasets are also scraped and sampled from available sources on the Internet, they may fail to represent all relevant user groups, and *historical bias* and *representation bias* can emerge. Additionally, if the benchmark datasets contain construct measures that fail to serve as a valid "proxy", *measurement bias* can emerge. Second, the composition of a red team can be biased and skew the evaluation results. Likewise, although machine learning researchers often have access to many statistical methods and models, they tend to select only a few results to report based on their personal preferences and available resources (Young, 2018). This selective reporting can create a "garden of forking paths" (Young, 2018), where different choices in the analysis process lead to significantly different and potentially incorrect results. This issue also arises in the development of LLMs. It underscores the importance of considering model uncertainty during the evaluation step to enhance the credibility and reliability of the models, especially given the growing scepticism and concerns about the potential harm from biased or incorrect outputs. This is crucial because the choice of performance metric, benchmark dataset and red teaming approach can all influence the evaluation results.

Phase 2: Customizing an LLM

LLM customization

After the base model is evaluated, education practitioners can tailor the model for their specific needs using various customization techniques (Figure 2). A popular technique for customizing an LLM is SFT (*supervised fine-tuning*), which refines the base model using a dataset to specialize the model in a particular domain or task (Liu et al., 2021; Zheng et al., 2023). For instance, *FineWeb-Edu* (Lozhkov et al., 2024) is an education-specific dataset (derived from the CommonCrawl dataset) comprising 1.3 trillion tokens for use in LLM customization. The resulting fine-tuned model retains the extensive knowledge embedded in the base model and additionally incorporates domain-specific information. In the education context, this method has been applied to improve automatic assessment scoring (Latif & Zhai, 2024), to support math tutors for remediation of students' mistakes (Wang et al., 2023), to assess personal qualities in college admission essays (Lira et al., 2023) and to reduce performance disparities in math problem skill tagging tasks across different languages (Kwak & Pardos, 2024).

Another potential technique for customizing an LLM is *preference tuning*, which is the process of adjusting a pre-trained model to better align it with specific preferences, priorities, or tastes. This can be accomplished using RLHF or direct preference optimization (DPO). DPO is a preference-tuning method inspired by reinforcement learning that is relatively simple, stable and computationally efficient; it outperforms commonly used methods such as proximal policy optimization (PPO) based RLHF in many cases (Rafailov et al., 2024). DPO leverages the relationship between the reward model and optimal policies, efficiently addressing the challenge of constrained reward optimization within a single policy training phase using human preference data. Both SFT and DPO techniques have been applied in education, for example, to create an intelligent question-answering system that is tailored to a specific introductory computer science course (Hicke et al., 2023).

When practitioners or researchers fine-tune a base model, their domain- or task-specific dataset and any human preference data they collect are vulnerable to both *historical bias* and *representation bias*. These datasets are often sourced from the Internet, smaller in size and focused narrowly on specialized tasks that reflect the characteristics of the domain. If this dataset mirrors skewed societal perspectives or inaccuracies or represents only a specific group of people, the fine-tuned model might adopt these biases, making it less

generalizable and more likely to make prejudiced decisions. Additionally, even if the dataset fairly represents the real world, *learning bias* can arise. In addition to amplifying undesirable biases in the training data, the model might overly adapt to the new dataset while updating parameters and forgetting some of the broader generalizations it had learned. This phenomenon is known as "catastrophic forgetting" (French, 1999), where domain-specific data overrides essential general knowledge. In the context of LLMs, Luo et al. (2023) conducted an empirical investigation and discovered that catastrophic forgetting is prevalent when fine-tuning LLMs such as Llama-7b and Alpaca-7B. Additionally, Zhai et al. (2023) found that fine-tuning multi-modal LLMs can lead to increased hallucinations. In the context of education, this could hypothetically mean that fine-tuning a base model on mathematics textbooks, for instance, could overly specialize the model in mathematics and cause it to give less helpful or even inaccurate responses in other subject areas. This would not necessarily be a problem if the model were used specifically for tasks like evaluating the correctness of mathematical answers or providing hints to students on mathematical problems. The issue would arise if the model is expected to also act holistically in a tutoring setting and handle a wide range of subjects requiring complex reasoning, as it might then fail to be an effective tutor.

The domain-specific dataset used for fine-tuning is typically collected either from a platform the LLM developers had already created or surveys, which can give rise to *measurement bias*. Measurement bias has been studied extensively in educational data, which commonly has a nested, multilevel structure because students are observed within classrooms, or each student might be given a different subset of questions for a standardized test (Jak et al., 2014). This type of measurement bias can be detected using structural equation modelling (SEM) with respect to different attributes, including student demographics, teacher demographics and classroom characteristics. Another example of measurement bias can arise from unexpected (and possibly unobserved) patterns in the data collection process. For example, Ogan et al. (2012) examined how an intelligent tutoring system (ITS) was used in classrooms in Latin America and found that many students worked collaboratively to solve problems, even though the system was designed for individual use. This can create measurement bias in ITS data that could be used for fine-tuning an LLM because high performance might be inaccurately attributed to individual students when they were actually collaborating.

If direct access to the fine-tuned LLM's internal parameters or embeddings is available, learning bias resulting from fine-tuning can be measured using the same embedding- and probability-based measures described above. Alternatively, extrinsic bias measures that systematically evaluate text generated by a fine-tuned model in response to specific prompts can be used (Delobelle et al., 2022). For this output evaluation, the token distribution between different social groups is compared using distribution metrics, classifier metrics or lexicon metrics. Distribution metrics compare the distribution of explicit or implicit mentions of social groups to a baseline distribution (Bommasani et al., 2023). These metrics compare differences in the percentage of predictions that exactly match the ground truth (ie, exact match) (Rajpurkar et al., 2016), or use co-occurrence measures (Bordia & Bowman, 2019) to detect variance in group representation. For example, the Perspective API (Google Jigsaw, 2024) measures toxicity by providing a toxicity probability for generated text. Sicilia and Alikhani (2023) suggested using Score Parity to assess how consistently a language model generates text based on certain attributes (eg, toxicity) across different protected attributes (eg, demographic groups). Lexicon-based metrics parse generated text at the word level, comparing words to pre-defined lists of harmful or biased terms, and assigning predefined bias scores to each word. Examples of lexicon-based metrics include HONEST (Nozza et al., 2022), which measures harmful words in generated text, and BOLD (Dhamala et al., 2021), which measures psycho-linguistic norms by assigning affective values (eg,

dominance, sadness) to words and calculating text-level norms as weighted averages; Gender Polarity (Dhamala et al., 2021) measures the frequency of gendered words in the generated text.

Another technique for LLM customization is *prompt customization* (also called "prompt tuning"). This method adapts pre-trained transformers to specific tasks by modifying the input prompts rather than changing the model's internal parameters (Liu et al., 2023). It leverages the inherent knowledge within pre-trained models to enhance its task-specific performance. Optimized prompt-tuning can be as effective as fine-tuning across models of various sizes and across different tasks (Liu et al., 2021). However, this technique can give rise to at least three types of *prompting bias*: majority label bias, recency bias (overemphasizing the importance of the latest information), and common token bias (Zhao et al., 2021). These can cause pre-trained LLMs to exhibit *representation bias* towards specific responses: for instance, if the final response prompt contains a negative label, it may influence the model to predict negative language. To measure prompting bias, Kotek et al. (2023) proposed a paradigm to test gender bias in LLMs by using a set of 15 prompts that contain stereotyping contexts to evaluate the susceptibility of a model.

Finally, a developer might use information retrieval in the LLM life cycle to generate responses grounded in information from relevant data sources, applying a customization technique called *retrieval-augmented generation* (RAG) (Lewis et al., 2020). This technique allows the model to refer to external information for generating responses using two primary types of retrieval: knowledge-based and API-based retrieval. Knowledge-based retrieval systems store the current context in a "vector store," an embedding space where users can query and find related content similar to the query. This ensures the LLM remains up-to-date and contextually relevant to specific downstream tasks. For example, a course-specific educational chatbot might use RAG to answer students' questions based on the official course materials (Hicke et al., 2023). API-based retrieval uses external databases, such as learning management systems (LMS) or student enrollment databases, to generate responses. This provides additional context, enhancing the quality of responses to be more personalized and relevant. However, combining LLMs with external databases to provide better context can introduce *representation bias* and *measurement bias* because of how contextual data is archived or integrated. Additionally, the retrieval system often uses ranking algorithms to sort the most similar contexts based on the user's query. These ranking systems can introduce *learning bias*, as they may favour certain types of content over others, possibly reducing the level of diversity in the set of retrieved documents.

Most LLM developers carefully review and monitor their base models and (to the extent possible) their customized models for the potential biases described above. They have formalized these checks into a set of *technical guardrails* for LLMs, which are critical frameworks and procedures aimed at promoting the ethical, secure, and accountable deployment of LLMs. These guardrails include content filters to prevent the generation of harmful or inappropriate content, usage monitoring to detect and mitigate misuse, and model tuning to reduce biases and enhance fairness. Additionally, technical guardrails may involve implementing privacy-preserving techniques to protect user data and incorporating explainability features to make the model's decisions more transparent (Attri, 2023). For example, Meta's Llama Guard (Inan et al., 2023) provides a holistic and thorough evaluation framework for responsible LLMs, while acknowledging its limitations, including that it focused on English, which can cause *representation bias* in other languages. Once again, as with any customization that attempts to measure and mitigate issues (eg, applying toxicity classification), the ML approach implemented within the technical guardrail is susceptible to *measurement bias* and *learning bias*.

Deployment (customized LLM → deployed system)

Once the customization is finalized and the quality of outputs is tested, the model can be deployed in various forms based on whether the model is solving NLU, NLG, or combined NLU-NLG tasks. When deploying a customized model such as Khanmigo (Khan Academy, n.d..) or Rori (Henkel et al., 2024), there can be a gap between the problem that the system was originally designed to address and the way it is used in practice, a so-called *deployment bias*. Deployment bias can be observed as a difference in application usage and performance across populations (Gallegos et al., 2024). For example, the study of students in South America using an ITS collaboratively, instead of individually as intended by the system designer, exemplifies the importance of deployment bias in authentic educational contexts (Ogan et al., 2012).

The deployment step in the LLM life cycle is particularly important because of the potential harm arising from human-computer interaction. When LLMs are used as "conversational agents" (Perez-Marin & Pascual-Nieto, 2011), they can "speak" in natural language, a primary mode of human communication. As a result, users might anthropomorphize these systems, viewing them as human-like, which can lead to overreliance or unsafe use. This presents a critical issue in education. For instance, if students over rely on an ITS because it appears adept at generating empathetic and expert responses, they may place undue trust in potentially unethical, unverified, or hallucinated information it generates. Students could be misled and engage in irresponsible academic practices. Another issue to consider with deployed LLM-based systems is how they might influence people's communication patterns. AI has been found to enhance communication efficiency and positive emotional expression, leading to closer and more cooperative interpersonal perceptions, but its use in conversations can be socially stigmatized and result in social harm (Hohenstein et al., 2023).

Finally, as LLMs are increasingly used to generate text, it is inevitable that text corpora used for training and customizing future models will include significant amounts of text generated by previous LLMs, rather than human authors. This may inadvertently amplify *historical and representation biases* that remained unaddressed in current LLMs (Wang et al., 2024), or biases that arise from who is predominantly generating text using LLMs and for what purposes. The continued induction of LLM-generated text into the population of all texts in the world will "pollute" datasets that represent human language, but it will also reflect the continuously evolving nature of language.

(Optional) LLM input to a larger model

Customized LLMs can be utilized by taking their outputs as inputs for more extensive machine learning frameworks. For example, an LLM can be used to assess students learning behaviours and status (eg, whether they experience confusion or have misconceptions that prevent them from solving a problem) (Li et al., 2024). Combining the NLU and NLG capabilities of LLMs can enable applications such as automated essay grading systems, which evaluate essays and give natural language feedback on aspects including statistical measures (eg, length and sentence complexity), stylistic elements (eg, syntax, grammar, and punctuation), and content quality (eg, accuracy, coherence, and key concept articulation) (Ramesh & Sanampudi, 2022). The prevalence of AI-generated content that could potentially influence student writing raises an emerging consideration about the ethics of AI plagiarism. For instance, LLMs can be used in an automated essay grading system to extract key features from student essays (NLU) and produce synthetic text (NLG) that can be used to check for originality against the student's work.

DISCUSSION

The life cycle of traditional machine learning applications, which focus on predicting predefined labels, is well understood (Suresh & Guttag, 2021). Biases are known to enter at various points in this life cycle, and methods to measure and mitigate these biases have been developed and tested, including in the context of education (Baker & Hawn, 2022; Kizilcec & Lee, 2022). However, with the increasing adoption of LLMs and other forms of generative AI in education, current evaluation approaches do not adequately address needs specific to supporting educational goals (Anthis et al., 2024; Denny et al., 2024; Yan et al., 2024). We introduce a framework for understanding the LLM life cycle, using domain-specific examples in education to highlight opportunities and challenges for incorporating NLU and NLG supports into education technology applications. We identify the potential sources of bias at each step of the LLM life cycle and discuss them in the context of education. This offers a framework for understanding where potential harms of LLMs might arise for students, teachers and other users of generative AI technology in education, which can guide approaches to bias measurement and mitigation.

Considering the important role of language in teaching, learning and other educational activities, LLMs will inevitably be part of AI-based educational decision support systems (AI-EDSS). For educational practitioners and policymakers, it is crucial to be well aware of the types of biases that can originate from various steps in the LLM life cycle. The life cycle perspective can offer them a heuristic for asking technology developers to explain each step to help them assess the risk of bias and potential harm. We have argued that measuring the biases in systems that use LLMs is more complex than in traditional ML, primarily because evaluating NLG is highly context-dependent; what constitutes good feedback on a homework assignment, for instance, can vary widely. Education technology developers can play a significant role in collecting and curating datasets for LLM evaluation and benchmarks, which can be combined with collections of educational content scraped from the Internet and filtered for quality, such as the *FineWeb-Edu* dataset (Lozhkov et al., 2024). To further tailor LLM-based systems for specific educational applications, participatory design methods to quickly prototype and collect feedback have been shown to work well in a holistic, evaluation-driven design approach, such as in the LearnLM project between Google and Arizona State University (Jurenka et al., 2024).

In developing the proposed framework, we reviewed how NLP research has approached bias and fairness issues, especially recent work with LLMs and we identified a few broader challenges. First, most evaluation protocols and objective functions are built with short feedback cycles, which is especially problematic in educational contexts that aim to support students' long-term growth as critical thinkers and problem solvers. This contrast is exemplified by the fact that most current evaluation methods examine just an isolated model output (single-turn), instead of an entire conversation (multi-turn), let alone evaluating how a conversation impacts future opportunities to demonstrate a deeper understanding of the topic.[3] Recent work using NLU to parse tutoring conversations to provide teachers with targeted feedback shows promise by moving towards a more holistic evaluation of multiple single-turn exchanges in a longer conversation (Demszky et al., 2023; Wang & Demszky, 2024). As the popularization of LLMs around 2023 happens to coincide with a broader movement in education to expand direct tutoring offerings (Loeb et al., 2023), it raises important questions about the effectiveness and responsible use of LLMs for on-demand tutoring, including as part of an ITS (D'Mello & Graesser, 2023).

Second, while this article focuses on understanding the sources of potential harms of LLMs in education, there is an important trade-off between those potential harms and the potential for substantial benefits that LLM-based education technologies can provide for students, teachers and their communities. We advocate for measuring and mitigate biases

in LLM-based systems, but at the same time, we must carefully consider the opportunity cost of delaying the adoption of systems that use LLMs due to concerns about bias. It is important to recognize that while LLMs can make errors or exhibit biases, they may actually be more accurate and consistent (and thus fairer) than humans (Kahneman et al., 2021), and their benefits can scale to much larger numbers. Many students receive little to no feedback on the assessments they complete, especially open-ended questions and longer pieces of writing. Providing access to instant feedback can significantly benefit students (Meyer et al., 2024), especially those in under-resourced environments, thereby potentially reducing achievement gaps and enhancing overall educational equity. We should not wait until LLMs are 100% free of bias, which is an unattainable goal to begin with (Anthis et al., 2024). We encourage education technology providers to responsibly explore innovations with LLM and large multi-modal model (LMM) technology to create more scalable engaging and effective learning experiences.

A third challenge we identified is the tensions between fairness and personalization. Most current research evaluates fairness as the uniformity of outputs across social groups or detects the prevalence of group-specific biases (Gallegos et al., 2024; Liang et al., 2023). This demands that students with comparable academic preparation and progress receive equivalent responses from LLMs, irrespective of their social identities, as in traditional machine learning contexts (Kizilcec & Lee, 2022). However, the complexity of social identities makes achieving absolute fairness an impractical goal and compromises are often needed to benefit from generally useful LLM applications at the cost of minor biases (Bell et al., 2023). There is a tension between two desirable but ostensibly incompatible properties of LLM applications: fairness, which demands that similar queries/students receive the same responses, and personalization, which encourages responses to be non-generic and tailored to students' needs. The pursuit of fairness might be at odds with the goal of personalized learning, which requires that LLM responses be customized to students' sociocultural backgrounds and individual needs, beyond accounting for their academic profiles (Anthis et al., 2024). Moreover, even if an LLM application could provide perfectly fair and personalized responses to students from different social backgrounds, such personalization might inadvertently widen existing gaps, as those already advantaged could benefit disproportionately from algorithmic support, thus challenging certain notions of fairness and equity (Dumont & Ready, 2023). Addressing these tensions presents a rich area for future research, which could clarify the philosophical and conceptual complexities of fairness in education when AI can provide highly personalized learning experiences, and establish an empirical foundation for evaluating and optimizing LLMs with fairness objectives in this context.

We note three limitations related to the proposed LLM life cycle framework. First, while we attempted to provide a holistic and complete account of the life cycle, there may be variations in the process of application development that are captured neither in Phase 1 nor 2. We still envision this framework to facilitate the identification and communication of biases, and to serve as a foundation to build on as the predominant life cycle may evolve. Second, the proposed framework does not provide guidance on the severity of harms that can result from various biases, because (a) any harms need to be evaluated in context, and (b) there have been too few studies to date that have demonstrated different kinds of biases in LLM-based applications in education. Third, while we highlight potential measures for each bias, this work does not comprehensively review bias measures and mitigation strategies for LLM-based applications in education.

We conclude with three recommendations for future research on LLMs in education. First, there is a need for education-focused benchmark datasets that better represent a broader range of sociodemographic groups across the world, especially considering that applications like Khanmigo are expected to be used by a diverse group of students and teachers (Gallegos et al., 2024). Additionally, there is a need for high-quality education

datasets for pre-training and fine-tuning models (Kwak & Pardos, 2024; Li et al., 2023; Lozhkov et al., 2024). Second, there is a need to develop a specific taxonomy of harms for LLMs in educational contexts that promote responsible use and highlight the perspectives of educators, students and their families. Current taxonomies tend to be domain-agnostic and developer-centred (eg, Weidinger et al. (2022)). Finally, there appears to be a significant opportunity to use high-quality human feedback from multi-turn scenarios to improve the efficacy and alignment of LLMs with educational objectives (Chung et al., 2024; Zhou et al., 2024), for instance, to refine them for specialized tasks such as math tutoring (Jurenka et al., 2024).

ACKNOWLEDGEMENTS

We thank Emma Harvey for comments on a previous version of the manuscript.

FUNDING INFORMATION

This work has been supported by funding from the Learning Engineering Virtual Institute (LEVI) and the Jacobs Foundation Research Fellowship.

CONFLICT OF INTEREST STATEMENT

There is no potential conflict of interest in this work.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ETHICS STATEMENT

There is no potential conflict of interest in this work.

ORCID

Jinsook Lee https://orcid.org/0000-0002-9957-1342
Yann Hicke https://orcid.org/0000-0001-7234-7001
Renzhe Yu https://orcid.org/0000-0002-2375-3537
Christopher Brooks https://orcid.org/0000-0003-0875-0204
René F. Kizilcec https://orcid.org/0000-0001-6283-5546

ENDNOTES

[1] https://huggingface.co/spaces/bigscience/license.

[2] In subsequent sections, we will be using the term "representation bias" to refer to both what Suresh and Guttag (Suresh & Guttag, 2021) refer to as representational bias (stereotypes, misrepresentation, toxic content, etc.) and also imbalances in training or fine-tuning datasets compared to a target population (eg an underrepresentation of women-authored texts), which is typically referred to as a representation bias. While these are distinct concepts, it is clear in a context which one is relevant, and so we opted for the simpler presentation by using one of the two phrases throughout.

[3] Single-turn evaluation is looking at {student utterance}–{teacher utterance} and evaluating how good the teacher utterance is based on the student utterance. It can be inside a conversation with many turns or on a question-answering platform with only one question followed by one answer. Multi-turn evaluation is looking at {student utterance}–{teacher utterance}–{student utterance}–{teacher utterance}–{student utterance}–{teacher utterance}–… and evaluating holistically the quality of the entire dialogue.

REFERENCES

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., & Penedo, G. (2023). The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., El Shafey, L., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., … Wu, Y. (2023). PaLM 2 technical report. *arXiv preprint arXiv:2305.10403*.

Anthis, J. R., Lum, K., Ekstrand, M., Feller, A., D'Amour, A., & Tan, C. (2024). The impossibility of fair LLMs. *arXiv e-prints, arXiv–2406*.

Attri. (2023). *A comprehensive guide: Everything you need to know about LLMs' guardrails*. https://attri.ai/blog/a-comprehensive-guide-everything-you-need-to-know-about-llms-guardrails

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., … Zhu, T. (2023). *Qwen technical report*. https://arxiv.org/abs/2309.16609

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *40*, 1052–1092.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671.

Bell, A., Bynum, L., Drushchak, N., Zakharchenko, T., Rosenblatt, L., & Stoyanovich, J. (2023). The possibility of fairness: Revisiting the impossibility theorem in practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 400–422). Association for Computing Machinery. https://doi.org/10.1145/3593013.3594007

Biever, C. (2023). ChatGPT broke the turing test-the race is on for new ways to assess AI. *Nature*, *619*(7971), 686–689.

BigScience Workshop, Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., … Wolf, T. (2023). *Bloom: A 176b-parameter open-access multilingual language model*.

Birhane, A., Prabhu, v., Han, S., Boddeti, V., & Luccioni, S. (2023). Into the LAION'S Den: Investigating hate in multimodal datasets. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. *36*, pp. 21268–21284). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/42f225509e8263e2043c9d834ccd9a2b-Paper-Datasets_and_Benchmarks.pdf

Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, *29*, 4349–4357.

Bommasani, R., Liang, P., & Lee, T. (2023). Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, *1525*(1), 140–146.

Bordia, S., & Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. In *North American chapter of the association for computational linguistics*. https://api.semanticscholar.org/CorpusID:102352788

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. *33*, pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. (2019). Understanding the origins of bias in word embeddings. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. *97*, p. Broader Impact of AI & ML Fairness). PMLR. http://proceedings.mlr.press/v97/brunet19a.html

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., … Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*(70), 1–53.

Coursera. (2023). *New products, tools, and features*. https://blog.coursera.org/new-products-tools-and-features-2023

Delobelle, P., Tokpo, E. K., Calders, T., & Berendt, B. (2022). Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American chapter of the association for computational linguistics* (pp. 1693–1706). Association for Computational Linguistics.

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2023). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*. https://doi.org/10.3102/01623737231169270

Denny, P., Gulwani, S., Heffernan, N. T., Käser, T., Moore, S., Rafferty, A. N., & Singla, A. (2024). Generative AI for education (GAIED): Advances, opportunities, and challenges. *arXiv preprint arXiv:2402.01580*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:52967399

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., & Gupta, R. (2021). BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 862–872). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445924

D'Mello, S. K., & Graesser, A. (2023). Intelligent tutoring systems: How computers achieve learning gains that rival human tutors. In *Handbook of educational psychology* (pp. 603–629). Routledge.

Dumont, H., & Ready, D. (2023). On the promise of personalized learning for educational equity. *npj Science of Learning*, *8*, 26. https://doi.org/10.1038/s41539-023-00174-x

Durrheim, K., Schuld, M., Mafunda, M., & Mazibuko, S. (2023). Using word embeddings to investigate cultural biases. *British Journal of Social Psychology*, *62*(1), 617–629.

edX Press. (n.d.). *edX Debuts Two AI-Powered Learning Assistants Built on ChatGPT*. https://press.edx.org/edx-debuts-two-ai-powered-learning-assistants-built-on-chatgpt

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–83. https://doi.org/10.1162/coli_a_00524

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., … Clark, J. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv: 2209.07858*.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., … Vinyals, O. (2024). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., … Kenealy, K. (2024). *Gemma: Open models based on Gemini research and technology. arXiv preprint arXiv:* 2403.08295.

Goldfarb-Tarrant, S., Marchant, R., Muñoz Sánchez, R., Pandya, M., & Lopez, A. (2021). Intrinsic bias metrics do not correlate with application bias. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long papers)* (pp. 1926–1940). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.150

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Google Jigsaw. (2024). Perspective API documentation. https://perspectiveapi.com/

Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., & Lee, A. (2024). Effective and scalable math support: Evidence on the impact of an AI- tutor on math achievement in Ghana. *arXiv preprint arXiv:2402.09809*.

Hicke, Y., Agarwal, A., Ma, Q., & Denny, P. (2023). Chata: Towards an intelligent question-answer teaching assistant using open-source LLMs. *arXiv preprint arXiv:2311.02775*.

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). Dialect prejudice predicts AI decisions about people's character, employability, and criminality. *arXiv preprint arXiv: 2403.00742*.

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., & Jung, M. F. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, *13*(1), 5487.

Hutiri, W. T., & Ding, A. Y. (2022). Bias in automated speaker recognition. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 230–247).

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., & Khabsa, M. (2023). Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*.

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 375–385). Online.

Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 31–39.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7B. *arXiv preprint arXiv: 2310.06825*.

Jurenka, I., Kunesch, M., McKee, K., Gillick, D., Zhu, S., Wiltberger, S., Phal, S. M., Hermann, K., Kasenberg, D., Bhoopchand, A., Anand, A., Pîslar, M., Chan, S., Wang, L., She, J., Mahmoudieh, P., Rysbek, A., Ko, W.-J., Huber, A., … Ibrahim, L. (2024). *Towards responsible development of generative AI for education: An evaluation-driven approach*. Google Technical Report. https://storage.googleapis.com/deepmind-media/LearnLM/LearnLM_paper.pdf

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Hachette UK.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). Chatgpt for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://www.sciencedirect.com/science/article/pii/S1041608023000195, https://doi.org/10.1016/j.lindif.2023.102274

Khan Academy. (n.d.). *Khan Academy Labs*. https://www.khanacademy.org/khan-labs

Kirk, H. R., Birhane, A., Vidgen, B., & Derczynski, L. (2022). Handling and presenting harmful text in NLP research. In *Findings of EMNLP*. Association for Computational Linguistics.

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In *The ethics of artificial intelligence in education* (pp. 174–202). Routledge.

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference* (pp. 12–24).

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Kwak, Y., & Pardos, Z. A. (2024). Bridging large language model disparities: Skill tagging of multilingual educational content. *British Journal of Educational Technology*, 1–19.

Latif, E., & Zhai, X. (2024). Fine-tuning chatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210.

LDNOOBW. (2023). *List of dirty, naughty, obscene, and otherwise bad words*. https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

Leiker, D., Finnigan, S., Gyllen, A. R., & Cukurova, M. (2023). Prototyping the use of large language models (LLMs) for adult learning content creation at scale. *arXiv preprint arXiv:2306.01815*.

Levin, N., Baker, R., Nasiar, N., Stephen, F., & Hutt, S. (2022). Evaluating gaming detector model robustness over time. In *Proceedings of the 15th International Conference on educational data mining, International Educational Data Mining Society*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. *33*, pp. 9459–9474). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Li, H., Li, C., Xing, W., Baral, S., & Heffernan, N. (2024). Automated feedback for student math responses based on multimodality and fine-tuning. In *Proceedings of the 14th learning analytics and knowledge conference* (pp. 763–770).

Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., & Lee, Y. T. (2023). Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., … Koreeda, Y. (2023). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Lin, J., Thomas, D. R., Han, F., Gupta, S., Tan, W., Nguyen, N. D., & Koedinger, K. R. (2023). Using large language models to provide explanatory feedback to human tutors. *arXiv preprint arXiv:2306.15498*.

Lira, B., Gardner, M., Quirk, A., Stone, C., Rao, A., Ungar, L., Hutt, S., Hickman, L., D'Mello, S. K., & Duckworth, A. L. (2023). Using artificial intelligence to assess personal qualities in college admissions. *Science Advances*, *9*(41), eadg9405.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35.

Liu, X., Ji, K., Fu, Y., Tam, W. L., Du, Z., Yang, Z., & Tang, J. (2021). *P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loeb, S., Novicoff, S., Pollard, C., Robinson, C., & White, S. (2023). *The effects of virtual tutoring on young readers: Results from a randomized controlled trial*. National Student Support Accelerator.

Lozhkov, A., Ben Allal, L., von Werra, L., & Wolf, T. (2024, May). *Fineweb-edu*. https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., & Zhang, Y. (2023). An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

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

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Nozza, D., Bianchi, F., Lauscher, A., & Hovy, D. (2022, May). Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali, & P. Buitelaar (Eds.), *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 26–34). Association for Computational Linguistics. https://aclanthology.org/2022.ltedi-1.4, https://doi.org/10.18653/v1/2022.ltedi-1.4

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, *45*(3), 487–501.

Ogan, A., Walker, E., Baker, R. S., Rebolledo Mendez, G., Jimenez Castro, M., Laurentino, T., & De Carvalho, A. (2012). Collaboration in cognitive tutor use in Latin America: Field study and design recommendations. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1381–1390).

OpenAI. (2023). *Gpt-4 technical report*. https://arxiv.org/abs/2303.08774

Pankiewicz, M., & Baker, R. S. (2024). Navigating compiler errors with AI assistance—A study of GPT hints in an introductory programming course. *arXiv preprint arXiv:2403.12737*.

Perez-Marin, D., & Pascual-Nieto, I. (2011). *Conversational agents and natural language interaction: Techniques and effective practices: Techniques and effective practices*. IGI Global.

Radford, A., & Narasimhan, K. (2018). *Improving language understanding by generative pre-training*. https://api.semanticscholar.org/CorpusID:49313245

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training. Open AI*. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. https://api.semanticscholar.org/CorpusID:160025533

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, *36*, 53728–53741.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527.

Sicilia, A., & Alikhani, M. (2023, July). Learning to generate equitable text in dialogue from biased training data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2898–2917). Association for Computational Linguistics. https://aclanthology.org/2023.acl-long.163, https://doi.org/10.18653/v1/2023.acl-long.163

Solaiman, I., & Dennison, C. (2021). Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, *34*, 5861–5873.

spamscanner. (2023). *Spam scanner: A node.js anti-spam, email filtering, and phishing prevention tool and service*. https://github.com/spamscanner/spamscanner

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM conference on equity and access in algorithms, mechanisms, and optimization* (Article 17, (pp. 1–9). Association for Computing Machinery. https://doi.org/10.1145/3465416.3483305

Swinger, N., De-Arteaga, M., Heffernan, N. T., IV, Leiserson, M. D., & Kalai, A. T. (2019). What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society* (pp. 305–311).

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *arXiv preprint arXiv:2311.14096*.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., … Kenealy, K. (2024). Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, A., Morgenstern, J., & Dickerson, J. P. (2024). Large language models cannot replace human participants because they cannot portray identity groups. *arXiv preprint arXiv:2402.01908*.

Wang, R. E., & Demszky, D. (2024). Edu-ConvoKit: An open-source library for education conversation data. *arXiv preprint arXiv:2402.05111*.

Wang, R. E., Zhang, Q., Robinson, C., Loeb, S., & Demszky, D. (2023). Step-by-step remediation of students' mathematical mistakes. *arXiv preprint arXiv:2310.10648*.

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., & Petrov, S. (2020). Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Weidinger, L., Mellor, J. F. J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). Ethical and social risks of harm from language models. *ArXiv*, *abs/2112.04359*. https://api.semanticscholar.org/CorpusID:244954639

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). Association for Computing Machinery. https://doi.org/10.1145/3531146.3533088

Weights & Biases. (2023). *Processing data for large language models*. https://wandb.ai/wandb_gen/llm-data-processing/reports/Processing-Data-for-Large-Language-Models--VmlldzozMDg4MTM2

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet?—A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, *4*, 654924.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112.

Young, C. (2018). Model uncertainty and the crisis in science. *Socius*, *4*, 2378023117737206.

Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y. J., & Ma, Y. (2023). Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K.-W. (2019). Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning* (pp. 12697–12706). PMLR.

Zheng, H., Shen, L., Tang, A., Luo, Y., Hu, H., Du, B., & Tao, D. (2023). Learn from model beyond fine-tuning: A survey. *arXiv preprint arXiv:2310.08184*.

Zheng, L., Niu, J., & Zhong, L. (2022). Effects of a learning analytics-based real-time feedback approach on knowledge elaboration, knowledge convergence, interactive relationships and group performance in CSCL. *British Journal of Educational Technology*, *53*(1), 130–149.

Zhou, Y., Zanette, A., Pan, J., Levine, S., & Kumar, A. (2024). Archer: Training language model agents via hierarchical multi-turn RL. *arXiv preprint arXiv:2402.19446*.

How to cite this article: Lee, J., Hicke, Y., Yu, R., Brooks, C., & Kizilcec, R. F. (2024). The life cycle of large language models in education: A framework for understanding sources of bias. *British Journal of Educational Technology*, *00*, 1–21. https://doi.org/10.1111/bjet.13505