Using large language models to facilitate academic work in psychological sciences

Aamir Sohail[1,2,3*†] and Lei Zhang[1,4,5]

[1] Centre for Human Brain Health, School of Psychology, University of Birmingham, Birmingham, UK
[2] Centre for Integrative Neuroscience and Neurodynamics, University of Reading, Reading, UK
[3] School of Psychology and Clinical Language Sciences, University of Reading, Reading, UK
[4] Institute for Mental Health, School of Psychology, University of Birmingham, Birmingham, UK
[5] Centre for Developmental Science, School of Psychology, University of Birmingham, Birmingham, UK

*Corresponding author: Aamir Sohail, Centre for Human Brain Health, School of Psychology, University of Birmingham, Birmingham, B15 2TT, UK

Aamir Sohail: Contact email (axs2210@bham.ac.uk)
ORCID: https://orcid.org/0009-0000-6584-4579

Lei Zhang: Contact email (l.zhang.13@bham.ac.uk)
ORCID: https://orcid.org/0000-0002-9586-595X

Statements and Declarations

Funding

A.S. is supported by an MRC AIM iCASE Grant (Ref: MR/W007002/1).

Competing interests

Both authors report no relevant financial or non-financial interests to disclose.

Author contributions

A.S. conceived the article, performed the literature search and drafted the work. A.S and L.Z. critically revised the work and approved the manuscript for publication.

Title

Using large language models to facilitate academic work in psychological sciences

Abstract

Large Language Models (LLMs) have significantly shaped working practices across a variety of fields including academia. Demonstrating a remarkable versatility, these models can generate responses to prompts with information in the form of text, documents, and images, show ability to summarise documents, perform literature searches, and even more, understand human behaviours. However, despite providing many clear benefits, barriers remain towards their integration into academic work. Ethical and practical concerns regarding their suitability for various tasks further complicate their appropriate use. Here, we summarise recent literature assessing the capacity of LLMs for different components of academic research and teaching, focusing on three key areas in the psychological sciences: education and assessment, academic writing, and simulating human behaviour. We discuss how LLMs can be used to aid each area, describe current challenges and good practices, and propose future directions. In doing so, we aim to increase the awareness and proper use of LLMs in various components of academic work, which will only feature more heavily over time.

Keywords: large language models (LLMs), academia, psychology, education, human behaviour, teaching

Introduction

Academics are expected to carry out teaching and research duties, having both a commitment to lecturing and grading student work, as well as designing and performing experiments, writing grant/funding applications, and publishing papers. This workload is often excessive, leading to long working hours and feelings of heightened anxiety and inefficiency (Barrett & Barrett, 2008). These burdens may be potentially alleviated by the recently developed large language models (LLMs; Vaswani et al., 2017). LLMs, a specific type of artificial neural networks that are pretrained on statistical relationships in language that ultimately generate a list of outcomes probabilistically representing the most suitable option in response to a given prompt (e.g., "Explain XYZ to first-year undergraduate students, are particularly suitable for specific tasks such as text summarization, knowledge retrieval, and cases where information can be concisely and accurately presented. Subsequently, these models can aid various components of academic work (Meyer et al., 2023), including in the psychological sciences (Abdurahman et al., 2023; Demszky et al., 2023), by summarising and revising text (Bekker, 2023), analysing and debugging computer code (Surameery & Shakor, 2023; Tian et al., 2023), and performing literature searches (Haman & Školník, 2023; Khraisha et al., 2024).

Teaching and academic writing are activities which particularly stand to benefit from the incorporation of LLMs, given that tasks in the psychological sciences heavily rely on text, verbal or written alike. Academics can use LLMs to freely generate content-relevant material (e.g., numerical cognition in infancy) and automate the grading of assessments, whilst students benefit from LLMs' utility as a knowledge base and ability to assist learning of practical skills including statistics and programming (e.g., general linear modelling in R; Wang et al., 2024). Similarly, LLMs also have significantly altered the writing process for academics, with its ability to propose templated articles, revise and re-word text, and perform literature searches in response to specific queries (Pinzolits, 2024). However, questions remain regarding their implementation for certain tasks, as LLMs often generate false information in response to specific prompts (Zhang et al., 2023) and false references when performing literature searches (Agrawal et al., 2024; Gao et al., 2023). Furthermore, students and academics, whilst benefitting from increased productivity, conversely face issues relating to

plagiarism (Hutson, 2024), critical thinking (Messeri & Crockett, 2024), and hinderances to the learning process (Yan et al., 2023).

Inherently rooted in the psychological sciences (particularly cognitive psychology), a common benchmark for understanding the capability of LLMs involves measuring the response to cognitive tasks and logic puzzles requiring 'human-like' reasoning (Huang & Chang, 2023). Early success in this domain (Kojima et al., 2022; Wei et al., 2022) prompted research towards using LLMs as proxies for human participants in behavioural experiments, potentially offering the ability to perform complex cognitive tasks more quickly, reliably and cheaply. Responding to behavioural tasks and other assessments submitted as prompts, LLMs are found to replicate classic economic, psycholinguistic, and social psychology experiments (Aher et al., 2023), ultimately demonstrating similarities with human cognition and behaviour (Agnew et al., 2024; Huijzer & Hill, 2023; Ke et al., 2024). However, others have noted the various biases inherent with LLMs, including differences between other measures of human decision-making and inference (Crockett & Messeri, 2023), and the inability to reflect more current or constantly changing societal views (Harding et al., 2023). It therefore currently remains unclear for academics in the psychological sciences to which extent LLMs can accurately represent human cognition, and the circumstances where they can accurately provide a substitution for human participants.

To reflect the state-of-the-art, this review summarizes the current development of research on LLMs in teaching, academic writing, and simulating human behaviour, in which we highlight the potential benefits and limitations for each. We then discuss the ethical considerations they present and suggest future directions in this rapidly evolving field.

Large language models in academic education

Psychology and related courses within higher education involve both theoretical and practical learning. Academics conceive and deliver concepts, theories, and empirical evidence for key topics in psychology, whereas students are expected to learn and portray critical insight towards those theories, and develop practical skills including statistics, experimental design, and programming. The underlying structure of LLMs make them highly suitable for aiding both theoretical and practical modes of learning, offering a clear benefit to both academics and students alike (Figure 1). Whilst the benefit for students is more apparent, teaching, at and above the undergraduate level, covers extensive amounts of conceptual information. As certain topics may initially be unfamiliar to the lecturer who will need to refresh their own subject knowledge, LLMs summarise complex topics at an appropriate level relevant for their teaching. LLMs can also be used to plan entire modules and how the content is delivered (Herft, 2023) by creating quizzes and assessments that test students' understanding of the material throughout the entire semester. This includes generating specific learning materials for those with learning difficulties (e.g., creating Concept Maps from conversations for dyslexic students) (D'Urso & Sciarrone, 2024) and translating materials into different languages for those whose primary language is not English (Lo, 2023). These elements are getting increasingly important considering equality, diversity, and inclusion in higher education (Prince & Francis, 2023). Ultimately, LLMs employed through chatbots such as ChatGPT benefit the teaching and learning process for both students and academics (Kasneci et al., 2023), improving student performance, motivation, organization and time management, and promotes a more effective and collaborative learning environment (Montenegro-Rueda et al., 2023; Yan et al., 2023).

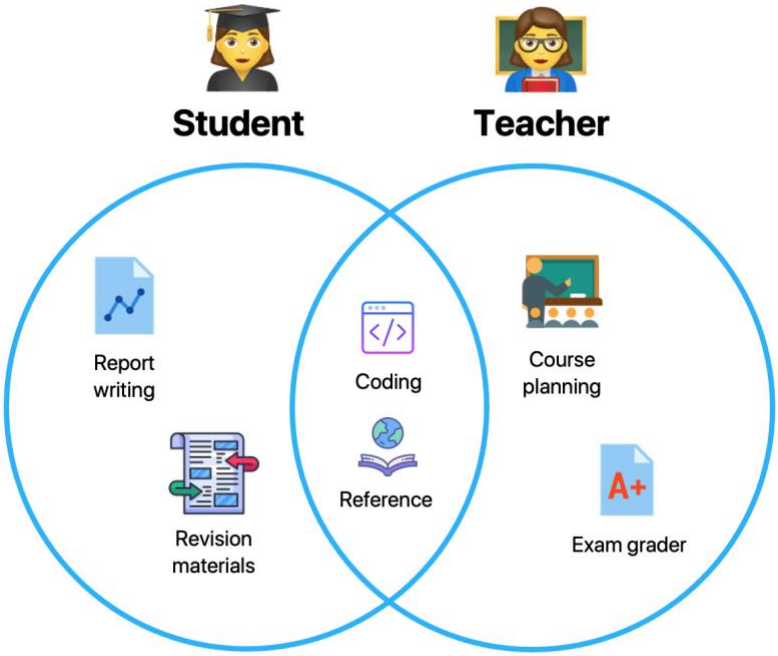

Figure 1. How academics and students can benefit from large language models (LLMs) in higher education. Demonstrating their versatility, large language models offer many benefits for both academics and students, most commonly by providing a knowledge base for key theories and concepts, and as a programming assistant. For students, LLMs can also assist with the revision process and at various stages of written coursework. Teachers can additionally benefit by using LLMs to plan courses and as an exam grader. Icons by Icons8.

From the students' perspective, LLMs can further benefit learning by generating educational materials such as reading comprehension tasks, interactive code explanations (Nam et al., 2024) and assessment questions (Shravya Bhat et al., 2022), and by improving student-based feedback of another's work (Jia et al., 2021). However, whether LLMs generally lead to an improvement in academic performance cannot be definitively stated, as there currently is a lack of empirically designed studies, particularly within the context of higher education (Kurtz et al., 2024).

The extent to which LLMs can bolster education is also dependent on the user's technical ability and personal attitudes. Certain academics report being reluctant to include LLMs as part of the learning process due to ethical concerns or unfamiliarity (Kiryakova & Angelova, 2023). Indeed, teachers in higher education also report confusion with adopting their curriculum accordingly given the prevalence of LLMs (Zhou et al., 2024). Conversely, many students also do not employ LLMs in their own learning, and if so, are not fully aware of its subtle nuances. Students new to programming – a common scenario in the psychological sciences - whilst aware that ChatGPT and other LLM-chatbots can be used to generate and fix code provided as prompts, may be under-educated in prompt engineering (Lin, 2024), the specific construction of prompts to receive a more suitable response (Avila-Chauvet et al., 2023). This is an important skill, as ChatGPT tends to be less capable in providing responses to programming questions if not well prompted (Kabir et al., 2023).

However, some have argued that an over-reliance on LLMs will have a negative influence on the skills and working practices accrued by students (Anders, 2023; Milano et al., 2023). Indeed, when using LLM tools to complete a programming project, students demonstrate practical progress but report hindered learning (Tanay et al., 2024), and a negative correlation has been observed between LLM reliance for programming tasks and performance on critical thinking assessments (Jošt et al.,

2024). By over-relying on the LLM to provide the solution, students may not think practically about the specific components of the code, resorting to simply copying and pasting generated code *ad nauseum*. We therefore suggest that students use LLMs in programming tasks (and similar tasks) in a scaffolding fashion – utilizing structures and pointers generated by LLMs as an "extra brain" yet independently evaluating and internalizing the actual solution, akin to the concept of zone of proximal development in development psychology (Vygotsky, 2012).

Yet, the lines regarding the appropriate use of LLMs in certain areas of education remain blurred. For example, in a programming class, should students be allowed to use code directly generated by an LLM? As employees are not restricted in the materials and resources available in their profession, some argue that universities should instead embrace LLMs and assess the efficacy in which students can use them to retrieve information and generate solutions (Koplin et al., 2023). Fully educating students on when (and when not) to use LLMs as part of their degree should therefore constitute a critical part of university-level education, avoiding the potential for an "unfair academic playing field", created by students unaware of the full capabilities of AI tools (Cotton et al., 2023), or those who choose not to use it due to ethical considerations (Grassini, 2023). In fact, a substantial number of universities worldwide have published student guidelines and guidance on using LLMs and generative artificial intelligence tools[1]. Meanwhile, online tools and platforms are publicly available (e.g., ChatGPT Detector, GPTZero) to detect work generated by LLMs to avoid overuse and misuse of LLMs in higher education.

Understanding the capabilities of LLMs also allows for academics, lecturers, and module convenors to set the appropriate examinations and assessments for their class. As these aim to measure subject knowledge, practical skills and critical thinking, abilities which can be replicated by LLMs to a degree, certain assessments in the psychological sciences may also need to be adjusted (Cotton et al., 2023; Rudolph et al., 2023). Attempts to prevent the use of LLMs for aiding assessments include employing AI-detectors for essays and reverting to oral presentations (Lemasters & Hurshman, 2024) and in-person written examinations. However, with the proven benefit in improving the learning process for certain areas, academics should remain open with students using LLMs in specific cases where the benefits in productivity can, but do not necessarily lead to, reduced learning. We ultimately advocate that academics are educated, well informed and develop a clear agenda before employing LLMs as a practical tool in their teaching.

Using large language models to aid academic writing

One of the more controversial issues regarding the use of LLMs within academia is their role with aiding the writing process. As LLMs can summarise, generate, and re-phrase text, journals have been quick to demonstrate their position on the matter, with some disallowing any LLM-generated text, and others requiring clear guidance as to which components of the research paper were influenced or generated (Curtis, 2023). Discerning to which extent LLMs should be used presents a difficult situation. Most would agree that entire paragraphs should not be written, re-written or paraphrased by LLMs; however, if, hypothetically, a human writer re-phrased a paragraph of academic text that coincidentally matched word-for-word an LLM-rephrased paragraph of the same text, should neither be used? Ethical dilemmas also exist on a smaller scale as the writing process naturally involves the repetition of others' work (this is particularly true for Methods sections in journal articles). Given that summarising the key results of a paper in a sentence or two can only contain a specific set of words, should LLMs be used to re-format a single sentence to avoid plagiarism? Some consider the

1 One of the example guidance is from the authors' affiliation: University of Birmingham (UK)'s Student guidance on using Generative Artificial Intelligence tools ethically for study. [retrieved on 09 July 2024]. https://intranet.birmingham.ac.uk/as/libraryservices/asc/student-guidance-gai.aspx

use of such programs even to re-structure single sentences as unacceptable in scientific research (Salvagno et al., 2023).

In a related but separate scenario, LLMs are often used to generate text intended for a research article or review paper from scratch by providing descriptions of scientific principles or an overview of a research topic. However, the underlying architecture of LLMs cautions against both uses. Answers provided by LLMs in response to open scientific questions can often be incorrect, or irrelevant (Hosseini et al., 2023), necessitating factual checking from the human user, whilst using LLMs to summarise research produces fabricated references (Day, 2023; Giray, 2023; Gravel et al., 2023), factually incorrect information (Han et al., 2023), and may be limited towards application and interpretation questions (Fergus et al., 2023). These false references may be entirely made-up, or legitimate articles with errors (Bhattacharyya et al., 2023), making it difficult for researchers to distinguish between the legitimate and illegitimate. Furthermore, using LLMs to summarise research areas has been found to generate inaccuracies compared to the published original work (Semrl et al., 2023). Paradoxically, however, the same study also demonstrated an ability to generate conclusions from provided abstracts indistinguishable from human-generated summaries, demonstrating an efficacy towards specific uses.

More recently, the performance of LLMs towards summarizing literature has improved due to the development of advanced models with larger training sets. Advanced and specialized search engines primarily implementing GPT-4 (e.g., SciSpace) can highlight relevant papers with fewer hallucinations and false references than earlier models. Whilst promising, these tools are still in their infancy and face several challenges, including hallucination and relevancy of papers to the prompt (Bolanos et al., 2024). One strategy is to restrict LLMs to aiding specific components of the literature review. For example, ChatGPT is able to generate research questions, suggest research terms and performs well in filtering and categorizing articles (Alshami et al., 2023), demonstrating that LLMs can rival human performance for certain review tasks including title/abstract screening, full-text review and data extraction (Khraisha et al., 2024). A hybrid model where LLMs identify relevant papers and themes, followed by the human-centered screening of relevant material, presents one such approach (Ye et al., 2024), both reducing errors and improving the accuracy of the literature review compared to a human-only workflow. Similar hybrid frameworks have been proposed for identifying elements in empirical papers, where LLMs present a time- and cost-effective approach whilst maintaining the accuracy observed in human reviewers (Uittenhove et al., 2024).

The versatility and extensive knowledge associated with LLMs stem from extensive pretraining on a wide, diverse corpus, the model subsequently acquiring a foundational grasp of both language and knowledge. Consequently, LLMs, including models trained upon enormous volumes of data, are still commonly not able to provide the domain-specific accuracy and precision in the information retrieved often essential for literature reviews (Susnjak et al., 2024). The accuracy of literature summarization for academic writing may therefore be improved by training LLMs on specific additional data, expanding the generalist knowledge with narrower domain-specific expertise. Whilst these “domain-specific” LLMs exist for a range of scientific disciplines, they are commonly associated with the medical sciences (Thirunavukarasu et al., 2023), potentially stemming from the high demand to summarise medical and biological information accurately and concisely for patient diagnoses and management. LLMs within this specific field are therefore specifically trained on large amounts of medical text. Reflecting a general trend of improvement in this space, performance on PubMedQA (Jin et al., 2019) a biomedical question answering dataset collated from PubMed abstracts, has improved over time with newer domain-specific models (Kamble & Alshikh, 2023; Singhal et al., 2023) displaying accuracies higher than base GPT-4 (Nori et al., 2023). Domain-specific LLMs vary in their subject expertise, from the broad topic of natural science (Xie et al., 2023), to niche areas including battery science (Zhao et al., 2024), and can be fine-tuned towards any scientific discipline. Certain workflows within this space also incorporate Retrieval-Augmented Generation

(RAG) (Lewis et al., 2021), which enhance the generation process by retrieving and incorporating relevant information associated with the user's input (Li et al., 2024), including relevant papers or keywords (Agarwal et al., 2024). Despite this, LLMs, whilst surpassing human capacity in certain components of the review process often lack the analytical depth and attention to detail that characterize human reviews (Tsai et al., 2024).

Using LLMs for proof-reading, editing, and shortening original text generated by the user are generally less contested within academia, as this occurs at the end of the creative process and leads to only minor changes from the original text. Some have likened this particular use of LLMs akin to asking a friend or colleague to proof-read a writing sample, which is unlikely to raise ethical concerns such as plagiarism that may arise under text summarization and generation (Meyer et al., 2023). Whilst early models were only able to process prompts in the form of text, more recently developed models can process entire documents, providing feedback on manuscripts within the order of seconds. Indeed, authors when presented with LLM-generated feedback on their published articles, find it helpful and more beneficial than feedback from some human reviewers (Liang et al., 2023). However, base models such as GPT-4 have been criticized for producing generic, non-meaningful comments, leading for tailored frameworks to be developed. Such frameworks typically levy multiple LLMs, assigning each LLM a specific task ultimately providing more meaningful and specific comments than the conventional single-model approach (D'Arcy et al., 2024; Gao et al., 2023). In any case, the accessibility of proof-reading and editing services through LLMs can additionally provide high-quality English language to non-native speakers and early-career researchers who would otherwise be placed at a disadvantage when submitting publications (Semrl et al., 2023). Indeed, in certain aspects, ChatGPT has proven to be more beneficial when compared to a paid English-editing service for academic editing (Kim, 2023). Proof-reading in the academic sphere can also be implemented to facilitate grant writing (Meyer et al., 2023) and to aid peer review (Hosseini & Horbach, 2023; Liu & Shah, 2023), allowing academics to focus more on new research (van Dis et al., 2023).

LLMs, whilst able to summarise and generate text as part of the academic writing process, currently demonstrate limitations in accuracy and legitimacy in certain domains, benefitting understanding and text analysis tasks more compared to literature review tasks (Wang et al., 2024). Therefore, whilst LLMs are able to rapidly generate a rapid, general overview of a subject, they currently fall short of being able to generate a literature review of the standards required in academia (Zimmermann et al., 2024). Assigning certain components of the workflow (e.g., identifying relevant papers) to LLMs can present a more time-effective approach whilst maintaining accuracy. Similar to other uses, benchmarking performance specific to searching and summarising scientific literature (Cai et al., 2024) is key for identifying their strengths and limitations within this space and supports the ongoing development of LLM workflows in scientific literature analysis.

Simulating human participants with large language models

Multiple fields of research including psychology, sociology, economics, and neuroscience utilise experiments to assess behaviours as part of their research methodology repertoire. However, despite its importance and usefulness, this process has several challenges and potential limitations, including high financial costs and data quality concerns (Chandler et al., 2014). Furthermore, human participants testing is also slowed by usually time-consuming ethical and practical components of the research process, requiring informed consent from participants, ethical approval, and additional requirements necessary for studying vulnerable groups. Some of the limitations and challenges associated with running behavioural experiments may therefore be avoided by employing artificial agents, with LLMs substituting for human participants.

Before diving into how LLMs can be useful in understanding human cognition, it is, first of all, important to unpack what "ability" is entailed in LLMs. One of the original motivations of developing LLMs and/or generative AI was to develop machines that could "think like humans" (Lake et al., 2017). The capacity of LLMs to do so stems from the numerous computational properties that allow these models to mimic and imitate human reasoning and inference (Aher et al., 2023). Certain models are further able to exhibit complex behaviour consistent with mentalistic inference (Strachan et al., 2024) and demonstrate similar heuristics and context-sensitive responses akin to loss aversion and effort reduction that are commonly observed in humans (Suri et al., 2024). LLMs are also more likely to succeed in some tasks and fail other tasks, just as human participants do (Dasgupta et al., 2023), leading for some researchers to state that the particular model tested could "pass as a valid subject" for some experiments (Binz & Schulz, 2023b). The appropriability for LLMs to do so is also improving over time, as important differences with human-like reasoning are prevalent in older models but disappear almost entirely in more recent ones (Yax et al., 2024), including a theory-of-mind (Strachan et al., 2024; Trott et al., 2023), demonstrating the importance of model size and complexity that could match the richness of human behaviours. Certain LLMs also demonstrate zero-shot learning (or generalisation), the ability to infer on data that the model have never seen in training, by accurately simulating human responses towards previously unseen cognitive tasks (Binz & Schulz, 2023a). Future research may seek to train LLMs on additional tasks, and novel tasks may eventually be tested on simulated cohorts, reducing time and financial costs in developing behavioural studies.

LLMs can also be experimentally induced into specific behavioural states through prompt engineering. For example, prompting LLMs with positive or negative components (e.g., adding the suffix 'This is very important to my career' or 'Perhaps this task is just beyond your skill set') has been found to affect the response generated (Li et al., 2023; X. Wang et al., 2024). This approach has subsequently been applied to understand psychopathology by inducing behavioural states observed among human cohorts with mental health conditions. By experimentally manipulating the level of 'anxiety' through anxiety-inducing and happiness-inducing scenarios, GPT-3.5 recreates performance characteristics observed in humans with high anxiety during a simple multi-armed bandit task, engaging in less exploitation and more exploration, and ultimately leading to worse behaviour (Coda-Forno et al., 2023). This and similar results have far-reaching implications for validating diagnostic measures and determining the efficacy of cognitive therapies, potentially in combination with computational and neuroimaging data of mental health conditions (Sohail & Zhang, 2024). Indeed, mindful-based interventions have been shown to reduce high levels of anxiety experimentally induced through traumatic narratives (Ben-Zion et al., 2024). As engineering positively themed prompts to LLMs shares similarities with delivering cognitive-based therapies in humans, prompts can be firstly fine-tuned in LLMs, with winning prompts subsequently tested in human patients. Early research has implemented such an LLM-informed treatment approach by generating dialogue systems based on Cognitive Behavioural Therapy (CBT) scenarios. Subsequently, patients report improved mood change and empathy to prompts generated by GPT-4, with no improvements to those generated by a dialogue model (Izumi et al., 2024).

Future studies could further utilize the same framework (i.e., first establish protocols in LLMs, then test it in humans) to investigate developmental psychopathology. Large language models display a pattern of increasing cognitive ability and rising language complexity in correspondence with child development, if prompted to do so (Milička et al., 2024). However, in the same study, task type, prompt type, and the choice of language model were all found to influence developmental patterns, demonstrating variability with this approach. Whilst LLMs offer a novel framework towards understanding human development, cognitive processes arising during childhood such as conceptual abstraction, should ideally be assessed using different tasks, and at multiple time points (Frank, 2023). Altogether, this recent and exciting field reflects similarities in computation between humans

and machines, with the potential for a computational psychiatric approach (Schulz & Dayan, 2020), informed by large language models.

Whilst LLMs can - in principle – be used as proxies for human participants, some have advised that this should only be done “when studying specific topics, when using specific tasks, at specific research stages, and when simulating specific samples” (Dillion et al., 2023), reflecting the differences in cognition and behaviour observed between humans and machines (Figure 2.). Albeit the important insights LLMs can offer in simulating human behaviours, LLMs have been shown to perform differently to human participants in many cognitive tasks, such as those necessitating directed exploration and causal reasoning (Binz & Schulz, 2023b), and during finitely-repeated economic games (Akata et al., 2023). Further differences between LLM and human task performance are illuminated through the "correct answer" effect, where questions probing political orientation, economic preference, judgement, and moral philosophy are answered with zero or near-zero variation (Park et al., 2023), ruling out the substitution of LLMs as human participants for certain tasks.

There are also questions into whether LLMs should be used at all in this manner concerning generalisability, as the training sets of LLMs are “HYPER-WEIRD”, overinfluenced by those from Western, Educated, Industrialized, Rich, Democratic (WEIRD) countries as well as those with attitudes that are Hegemonic, Young, and Publicly ExpRessed (Crockett & Messeri, 2023). Consequently, these models may lack sufficient diversity in their responses to accurately represent a representative population sample (Wang et al., 2024). ChatGPT, for example, demonstrates gender (Ghosh & Caliskan, 2023), cultural (Cao et al., 2023), and political (Hartmann et al., 2023) biases in its responses, and shows significantly less variance compared to human participants across a range of self-report measures spanning various psychological domains, such as personality, cognition, political orientation, and emotions (Atari et al., 2023). Indeed, GPT-4 has been described as having both increased honesty and humility and demonstrating masculine and anxious traits (Barua et al., 2024). Substituting participants for LLMs could therefore propagate the over-sampling of a specific sub-population, the antithesis of psychological research which is often to obtain samples from and to make inferences towards diverse populations. In addition, LLMs also demonstrate – relative to human responses - increased susceptibility to biases such as irrationality (Alsagheer et al., 2024), and response inconsistency (Macmillan-Scott & Musolesi, 2024), and are influenced by unknowingly biased features of model inputs (Turpin et al., 2023), including language (Goli & Singh, 2024).

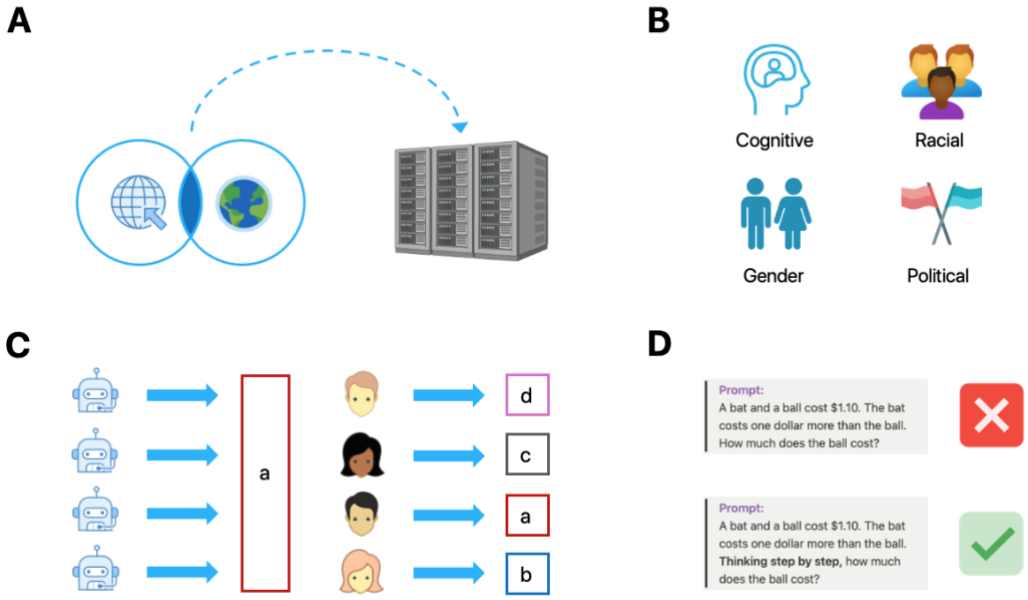

Figure 2. Considerations of employing large language models (LLMs) as proxies for human participants.
(A) Models are trained upon large quantities of online data influenced by those with access to the internet, unrepresentative of the human population. **(B)** LLMs demonstrate several biases including cognitive, racial, gender and political inclinations in their responses to specific prompts. **(C)** In response to questions probing political orientation, economic preference, and moral philosophy, human cohorts demonstrate significant response variability whereas LLMs demonstrate near-zero variation, a phenomenon dubbed the 'correct answer effect'. **(D)** Prompt engineering significantly influences the response provided by LLMs, whilst having little effect on human-based reasoning (Yax et al., 2024). Depicted is the 'Chain-of-Thought' prompt engineering strategy which improves LLM-based reasoning by breaking down the response into discrete steps. Icons by Icons8.

Despite these concerns, LLMs provide a tangible benefit as proxies for human participants for specific experimental designs not susceptible to cognitive or variational biases. Looking forward, this promising field should further identify the similarities and differences between LLMs and human behaviour by developing testable and ethologically meaningful benchmarks (Coda-Forno et al., 2024; Gandhi et al., 2023; Huang et al., 2024; Momennejad et al., 2023), frameworks guiding experimenters whether to integrate LLM-generated data into their research pipeline (Trott, 2024), and prompt datasets for mitigating against cognitive biases (Echterhoff et al., 2024). Furthermore, making publicly available articles, tutorials, and notebooks detailing how the lay-psychology researcher can substitute LLMs for human participants (Hussain et al., 2023) will make this often technically difficult research more accessible within the psychological sciences. It is worth noting several challenges towards future development in this vein. Developing benchmarks for social biases are often subjective and context-dependent, and are not detected by automated benchmarks and objective metrics such as accuracy (Aoyagui et al., 2024). Furthermore, the varying accuracy observed between different models and human responses with defining broad concepts has led for some to necessitate the definition of more specified cognitive measures (Almeida et al., 2024). In the face of these challenges, if appropriately used, LLMs have the potential to significantly change how academic experiments are conducted (Huijzer & Hill, 2023).

Conclusion

LLMs contest a highly debated area of academic research including the psychological sciences. Whilst it is not quite the 'academic panacea' (Quintans-Júnior et al., 2023) some have made it to be, LLMs constitute an integral part of the academic workflow for an increasing number. Academics currently use LLMs to write essays and talks, summarize literature, draft and improve papers, identify research gaps, write computer code and perform statistical analyses. As time progresses, this capability will only increase, evolving to the point that LLMs are expected to design experiments, write and complete manuscripts, conduct peer review and support editorial decisions to accept or reject manuscripts. Furthermore, domain-specific LLMs will further increase academic performance and productivity within specific fields. For those with little experience, a progressive adoption model, where LLMs are gradually incorporated into academic work (Kurtz et al., 2024) is recommended. Whilst managing a balance between efficiency and legitimacy of both teaching and research will be a difficult challenge, we nevertheless advocate for LLMs to be openly endorsed by academics in psychology and beyond.

References

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2023). *Perils and Opportunities in Using Large Language Models in Psychological Research*. https://doi.org/10.31234/osf.io/d695y

Agarwal, S., Laradji, I. H., Charlin, L., & Pal, C. (2024). *LitLLM: A Toolkit for Scientific Literature Review* (arXiv:2402.01788). arXiv. https://doi.org/10.48550/arXiv.2402.01788

Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024). The illusion of artificial inclusion. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3613904.3642703

Agrawal, A., Suzgun, M., Mackey, L., & Kalai, A. T. (2024). *Do Language Models Know When They're Hallucinating References?* (arXiv:2305.18248). arXiv. https://doi.org/10.48550/arXiv.2305.18248

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). *Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies* (arXiv:2208.10264). arXiv. https://doi.org/10.48550/arXiv.2208.10264

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., & Schulz, E. (2023). *Playing repeated games with Large Language Models* (arXiv:2305.16867). arXiv. https://doi.org/10.48550/arXiv.2305.16867

Almeida, G. F. C. F., Nunes, J. L., Engelmann, N., Wiegmann, A., & Araújo, M. de. (2024). Exploring the psychology of LLMs' moral and legal reasoning. *Artificial Intelligence*, *333*, 104145. https://doi.org/10.1016/j.artint.2024.104145

Alsagheer, D., Karanjai, R., Diallo, N., Shi, W., Lu, Y., Beydoun, S., & Zhang, Q. (2024). *Comparing Rationality Between Large Language Models and Humans: Insights and Open Questions* (arXiv:2403.09798). arXiv. https://doi.org/10.48550/arXiv.2403.09798

Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. E. E., & Zayed, T. (2023). Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. *Systems*, *11*(7), Article 7. https://doi.org/10.3390/systems11070351

Anders, B. A. (2023). Is using ChatGPT cheating, plagiarism, both, neither, or forward thinking? *Patterns*, *4*(3), 100694. https://doi.org/10.1016/j.patter.2023.100694

Aoyagui, P. A., Ferguson, S., & Kuzminykh, A. (2024). *Exploring Subjectivity for more Human-Centric Assessment of Social Biases in Large Language Models* (arXiv:2405.11048). arXiv. https://doi.org/10.48550/arXiv.2405.11048

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). *Which Humans?* OSF. https://doi.org/10.31234/osf.io/5b26t

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.903077

Avila-Chauvet, L., Mejía, D., & Acosta Quiroz, C. O. (2023). *Chatgpt as a Support Tool for Online Behavioral Task Programming* (SSRN Scholarly Paper 4329020). https://papers.ssrn.com/abstract=4329020

Barrett, L. C., & Barrett, P. S. (2008). *The management of academic workloads: Full report on findings*. https://salford-repository.worktribe.com/output/1472770

Barua, A., Brase, G., Dong, K., Hitzler, P., & Vasserman, E. (2024). *On the Psychology of GPT-4: Moderately anxious, slightly masculine, honest, and humble* (arXiv:2402.01777). arXiv. https://doi.org/10.48550/arXiv.2402.01777

Bekker, M. (2023). *Large Language Models and Academic Writing: Five tiers of engagement*. OSF Preprints. https://doi.org/10.31219/osf.io/63vcu

Ben-Zion, Z., Witte, K., Jagadish, A. K., Duek, O., Harpaz-Rotem, I., Khorsandian, M.-C., Burrer, A., Seifritz, E., Homan, P., Schulz, E., & Spiller, T. R. (2024). *"Chat-GPT on the Couch": Assessing and Alleviating State Anxiety in Large Language Models*. OSF. https://doi.org/10.31234/osf.io/j7fwb

Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., & Miller, L. E. (2023). High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Cureus*, *15*(5), e39238. https://doi.org/10.7759/cureus.39238

Binz, M., & Schulz, E. (2023a). *Turning large language models into cognitive models* (arXiv:2306.03917). arXiv. https://doi.org/10.48550/arXiv.2306.03917

Binz, M., & Schulz, E. (2023b). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Bolanos, F., Salatino, A., Osborne, F., & Motta, E. (2024). *Artificial Intelligence for Literature Reviews: Opportunities and Challenges* (arXiv:2402.08565). arXiv. https://doi.org/10.48550/arXiv.2402.08565

Cai, H., Cai, X., Chang, J., Li, S., Yao, L., Wang, C., Gao, Z., Wang, H., Li, Y., Lin, M., Yang, S., Wang, J., Yin, Y., Li, Y., Zhang, L., & Ke, G. (2024). *SciAssess: Benchmarking LLM Proficiency in Scientific Literature Analysis* (arXiv:2403.01976). arXiv. https://doi.org/10.48550/arXiv.2403.01976

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). *Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study* (arXiv:2303.17466). arXiv. https://doi.org/10.48550/arXiv.2303.17466

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. https://doi.org/10.3758/s13428-013-0365-7

Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?* (arXiv:2307.09009). arXiv. https://doi.org/10.48550/arXiv.2307.09009

Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., & Schulz, E. (2023). *Inducing anxiety in large language models increases exploration and bias* (arXiv:2304.11111). arXiv. https://doi.org/10.48550/arXiv.2304.11111

Coda-Forno, J., Binz, M., Wang, J. X., & Schulz, E. (2024). *CogBench: A large language model walks into a psychology lab* (arXiv:2402.18225). arXiv. https://doi.org/10.48550/arXiv.2402.18225

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *0*(0), 1–12. https://doi.org/10.1080/14703297.2023.2190148

Crockett, M., & Messeri, L. (2023). *Should large language models replace human participants?* PsyArXiv. https://doi.org/10.31234/osf.io/4zdx9

Curtis, N. (2023). To ChatGPT or not to ChatGPT? The Impact of Artificial Intelligence on Academic Publishing. *The Pediatric Infectious Disease Journal*, *42*(4), 275. https://doi.org/10.1097/INF.0000000000003852

D'Arcy, M., Hope, T., Birnbaum, L., & Downey, D. (2024). *MARG: Multi-Agent Review Generation for Scientific Papers* (arXiv:2401.04259). arXiv. https://doi.org/10.48550/arXiv.2401.04259

D'Urso, S., & Sciarrone, F. (2024). AI4LA: An Intelligent Chatbot for Supporting Students with Dyslexia, Based on Generative AI. In A. Sifaleras & F. Lin (Eds.), *Generative Intelligence and Intelligent Tutoring Systems* (pp. 369–377). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-63028-6_31

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2023). *Language models show human-like content effects on reasoning tasks* (arXiv:2207.07051). arXiv. https://doi.org/10.48550/arXiv.2207.07051

Day, T. (2023). A Preliminary Investigation of Fake Peer-Reviewed Citations and References Generated by ChatGPT. *The Professional Geographer*, *0*(0), 1–4. https://doi.org/10.1080/00330124.2023.2190373

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701. https://doi.org/10.1038/s44159-023-00241-5

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024). *Cognitive Bias in High-Stakes Decision-Making with LLMs* (arXiv:2403.00811). arXiv. https://doi.org/10.48550/arXiv.2403.00811

Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating Academic Answers Generated Using ChatGPT. *Journal of Chemical Education*, *100*(4), 1672–1675. https://doi.org/10.1021/acs.jchemed.3c00087

Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, *2*(8), 451–452. https://doi.org/10.1038/s44159-023-00211-x

Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. D. (2023). *Understanding Social Reasoning in Language Models with Language Models* (arXiv:2306.15448). arXiv. https://doi.org/10.48550/arXiv.2306.15448

Gao, T., Yen, H., Yu, J., & Chen, D. (2023). *Enabling Large Language Models to Generate Text with Citations* (arXiv:2305.14627). arXiv. https://doi.org/10.48550/arXiv.2305.14627

548 Gao, Z., Brantley, K., & Joachims, T. (2024). *Reviewer2: Optimizing Review Generation Through*
549 *Prompt Generation* (arXiv:2402.10886). arXiv. https://doi.org/10.48550/arXiv.2402.10886

550 Ghosh, S., & Caliskan, A. (2023). *ChatGPT Perpetuates Gender Bias in Machine Translation and*
551 *Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource*
552 *Languages* (arXiv:2305.10510). arXiv. https://doi.org/10.48550/arXiv.2305.10510

553 Giray, L. (2023). ChatGPT References Unveiled: Distinguishing the Reliable from the Fake. *Internet*
554 *Reference Services Quarterly*, *0*(0), 1–10. https://doi.org/10.1080/10875301.2023.2265369

555 Goli, A., & Singh, A. (2024). Frontiers: Can Large Language Models Capture Human Preferences?
556 *Marketing Science*. https://doi.org/10.1287/mksc.2023.0306

557 Grassini, S. (2023). Shaping the Future of Education: Exploring the Potential and Consequences of AI
558 and ChatGPT in Educational Settings. *Education Sciences*, *13*(7), Article 7.
559 https://doi.org/10.3390/educsci13070692

560 Gravel, J., D'Amours-Gravel, M., & Osmanlliu, E. (2023). Learning to Fake It: Limited Responses and
561 Fabricated References Provided by ChatGPT for Medical Questions. *Mayo Clinic Proceedings:*
562 *Digital Health*, *1*(3), 226–234. https://doi.org/10.1016/j.mcpdig.2023.05.004

563 Haman, M., & Školník, M. (2023). Using ChatGPT to conduct a literature review. *Accountability in*
564 *Research*, *0*(0), 1–3. https://doi.org/10.1080/08989621.2023.2185514

565 Han, Z., Battaglia, F., Udaiyar, A., Fooks, A., & Terlecky, S. R. (2023). *An Explorative Assessment of*
566 *ChatGPT as an Aid in Medical Education: Use it with Caution* (p. 2023.02.13.23285879).
567 medRxiv. https://doi.org/10.1101/2023.02.13.23285879

568 Harding, J., D'Alessandro, W., Laskowski, N. G., & Long, R. (2023). AI language models cannot replace
569 human research participants. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01725-x

570 Hartmann, J., Schwenzow, J., & Witte, M. (2023). *The political ideology of conversational AI:*
571 *Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation*
572 (arXiv:2301.01768). arXiv. https://doi.org/10.48550/arXiv.2301.01768

573 Herft, A. (2023). *A Teacher's Prompt Guide to ChatGPT aligned with 'What Works Best'.pdf*. Google
574 Docs.
575 https://drive.google.com/file/d/15qAxnUzOwAPwHzoaKBJd8FAgiOZYcIxq/view?usp=embed
576 _facebook

577 Horton, J. J. (2023). *Large Language Models as Simulated Economic Agents: What Can We Learn*
578 *from Homo Silicus?* (Working Paper 31122). National Bureau of Economic Research.
579 https://doi.org/10.3386/w31122

580 Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias?
581 Considerations and recommendations for use of ChatGPT and other large language models
582 in scholarly peer review. *Research Integrity and Peer Review*, *8*(1), 4.
583 https://doi.org/10.1186/s41073-023-00133-5

584 Hosseini, M., Rasmussen, L. M., & Resnik, D. B. (2023). Using AI to write scholarly publications.
585 *Accountability in Research*, *0*(0), 1–9. https://doi.org/10.1080/08989621.2023.2168535

Huang, J., & Chang, K. C.-C. (2023). *Towards Reasoning in Large Language Models: A Survey* (arXiv:2212.10403). arXiv. https://doi.org/10.48550/arXiv.2212.10403

Huang, J., Wang, W., Li, E. J., Lam, M. H., Ren, S., Yuan, Y., Jiao, W., Tu, Z., & Lyu, M. R. (2024). *Who is ChatGPT? Benchmarking LLMs' Psychological Portrayal Using PsychoBench* (arXiv:2310.01386). arXiv. https://doi.org/10.48550/arXiv.2310.01386

Huijzer, R., & Hill, Y. (2023). *Large Language Models Show Human Behavior*. https://doi.org/10.31234/osf.io/munc9

Hussain, Z., Binz, M., Mata, R., & Wulff, D. U. (2023). *A tutorial on open-source large language models for behavioral science*. https://doi.org/10.31234/osf.io/f7stn

Hutson, J. (2024). Rethinking Plagiarism in the Era of Generative AI. *Journal of Intelligent Communication*, *4*(1), Article 1. https://doi.org/10.54963/jic.v4i1.220

Izumi, K., Tanaka, H., Shidara, K., Adachi, H., Kanayama, D., Kudo, T., & Nakamura, S. (2024). *Response Generation for Cognitive Behavioral Therapy with Large Language Models: Comparative Study with Socratic Questioning* (arXiv:2401.15966). arXiv. https://doi.org/10.48550/arXiv.2401.15966

Jia, Q., Cui, J., Xiao, Y., Liu, C., Rashid, P., & Gehringer, E. F. (2021). *ALL-IN-ONE: Multi-Task Learning BERT models for Evaluating Peer Assessments* (arXiv:2110.03895). arXiv. https://doi.org/10.48550/arXiv.2110.03895

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). *PubMedQA: A Dataset for Biomedical Research Question Answering* (arXiv:1909.06146). arXiv. https://doi.org/10.48550/arXiv.1909.06146

Jošt, G., Taneski, V., & Karakatič, S. (2024). The Impact of Large Language Models on Programming Education and Student Learning Outcomes. *Applied Sciences*, *14*(10), Article 10. https://doi.org/10.3390/app14104115

Kabir, S., Udo-Imeh, D. N., Kou, B., & Zhang, T. (2023). *Who Answers It Better? An In-Depth Analysis of ChatGPT and Stack Overflow Answers to Software Engineering Questions* (arXiv:2308.02312). arXiv. https://doi.org/10.48550/arXiv.2308.02312

Kamble, K., & Alshikh, W. (2023). *Palmyra-Med: Instruction-Based Fine-Tuning of LLMs Enhancing Medical Domain Performance*. https://doi.org/10.13140/RG.2.2.30939.75046

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://doi.org/10.1016/j.lindif.2023.102274

Ke, L., Tong, S., Cheng, P., & Peng, K. (2024). *Exploring the Frontiers of LLMs in Psychological Applications: A Comprehensive Review* (arXiv:2401.01519). arXiv. https://doi.org/10.48550/arXiv.2401.01519

Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., & Hadfield, K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and

extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods*. https://doi.org/10.1002/jrsm.1715

Kim, S.-G. (2023). Using ChatGPT for language editing in scientific articles. *Maxillofacial Plastic and Reconstructive Surgery*, *45*(1), 13. https://doi.org/10.1186/s40902-023-00381-x

Kiryakova, G., & Angelova, N. (2023). ChatGPT—A Challenging Tool for the University Professors in Their Teaching Practice. *Education Sciences*, *13*(10), Article 10. https://doi.org/10.3390/educsci13101056

Kojima, T., Gu, S. (Shane), Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.

Koplin, J., Sparrow, R., Hatherley, J., Rivers, N., & Williams, I. (2023, May 15). *Tailoring university assessment in the age of ChatGPT*. Monash Lens. https://lens.monash.edu/@politics-society/2023/05/15/1385696?slug=tailoring-university-assessment-in-the-age-of-chatgpt

Kurtz, G., Amzalag, M., Shaked, N., Zaguri, Y., Kohen-Vacs, D., Gal, E., Zailer, G., & Barak-Medina, E. (2024). Strategies for Integrating Generative AI into Higher Education: Navigating Challenges and Leveraging Opportunities. *Education Sciences*, *14*(5), Article 5. https://doi.org/10.3390/educsci14050503

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Lemasters, R., & Hurshman, C. (2024). A shift towards oration: Teaching philosophy in the age of large language models. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00455-0

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* (arXiv:2005.11401). arXiv. https://doi.org/10.48550/arXiv.2005.11401

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., & Xie, X. (2023). Large Language Models Understand and Can be Enhanced by Emotional Stimuli. In *arXiv e-prints*. https://doi.org/10.48550/arXiv.2307.11760

Li, Y., Chen, L., Liu, A., Yu, K., & Wen, L. (2024). *ChatCite: LLM Agent with Human Workflow Guidance for Comparative Literature Summary* (arXiv:2403.02574). arXiv. https://doi.org/10.48550/arXiv.2403.02574

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., McFarland, D., & Zou, J. (2023). *Can large language models provide useful feedback on research papers? A large-scale empirical analysis* (arXiv:2310.01783). arXiv. https://doi.org/10.48550/arXiv.2310.01783

Lin, Z. (2024). How to write effective prompts for large language models. *Nature Human Behaviour*, *8*(4), 611–615. https://doi.org/10.1038/s41562-024-01847-2

Liu, R., & Shah, N. B. (2023). *ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing* (arXiv:2306.00622). arXiv. https://doi.org/10.48550/arXiv.2306.00622

Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences*, *13*(4), Article 4. https://doi.org/10.3390/educsci13040410

Macmillan-Scott, O., & Musolesi, M. (2024). (Ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, *11*(6), 240255. https://doi.org/10.1098/rsos.240255

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O'Connor, K., Li, R., Peng, P.-C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., & Moore, J. H. (2023). ChatGPT and large language models in academia: Opportunities and challenges. *BioData Mining*, *16*(1), 20. https://doi.org/10.1186/s13040-023-00339-9

Milička, J., Marklová, A., VanSlambrouck, K., Pospíšilová, E., Šimsová, J., Harvan, S., & Drobil, O. (2024). Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLOS ONE*, *19*(3), e0298522. https://doi.org/10.1371/journal.pone.0298522

Momennejad, I., Hasanbeig, H., Vieira, F., Sharma, H., Ness, R. O., Jojic, N., Palangi, H., & Larson, J. (2023). *Evaluating Cognitive Maps and Planning in Large Language Models with CogEval* (arXiv:2309.15129). arXiv. https://doi.org/10.48550/arXiv.2309.15129

Montenegro-Rueda, M., Fernández-Cerero, J., Fernández-Batanero, J. M., & López-Meneses, E. (2023). Impact of the Implementation of ChatGPT in Education: A Systematic Review. *Computers*, *12*(8), Article 8. https://doi.org/10.3390/computers12080153

Nam, D., Macvean, A., Hellendoorn, V., Vasilescu, B., & Myers, B. (2024). Using an LLM to Help With Code Understanding. *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 1–13. https://doi.org/10.1145/3597503.3639187

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). *Capabilities of GPT-4 on Medical Challenge Problems* (arXiv:2303.13375). arXiv. https://doi.org/10.48550/arXiv.2303.13375

Park, P. S., Schoenegger, P., & Zhu, C. (2023). *Diminished Diversity-of-Thought in a Standard Large Language Model* (arXiv:2302.07267). arXiv. https://doi.org/10.48550/arXiv.2302.07267

Pinzolits, R. (2024). AI in academia: An overview of selected tools and their areas of application. *MAP Education and Humanities*, *4*, 37–50. https://doi.org/10.53880/2744-2373.2023.4.37

Plevris, V., Papazafeiropoulos, G., & Rios, A. J. (2023). *Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard* (arXiv:2305.18618). arXiv. https://doi.org/10.48550/arXiv.2305.18618

Prince, L. R., & Francis, S. E. (2023). Barriers to equality, diversity and inclusion in research and academia stubbornly persist. So, what are we doing about it? *Disease Models & Mechanisms*, *16*(7), dmm050048. https://doi.org/10.1242/dmm.050048

Quintans-Júnior, L. J., Gurgel, R. Q., Araújo, A. A. de S., Correia, D., & Martins-Filho, P. R. (2023). ChatGPT: The new panacea of the academic world. *Revista Da Sociedade Brasileira de Medicina Tropical*, *56*, e0060. https://doi.org/10.1590/0037-8682-0060-2023

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1), Article 1. https://doi.org/10.37074/jalt.2023.6.1.9

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? *Critical Care*, *27*(1), 75. https://doi.org/10.1186/s13054-023-04380-2

Schulz, E., & Dayan, P. (2020). Computational Psychiatry for Computers. *iScience*, *23*(12), 101772. https://doi.org/10.1016/j.isci.2020.101772

Semrl, N., Feigl, S., Taumberger, N., Bracic, T., Fluhr, H., Blockeel, C., & Kollmann, M. (2023). AI language models in human reproduction research: Exploring ChatGPT's potential to assist academic writing. *Human Reproduction*, dead207. https://doi.org/10.1093/humrep/dead207

Shravya Bhat, Nguyen, H., Moore, S., Stamper, J., Sakr, M., & Nyberg, E. (2022). *Towards Automated Generation and Evaluation of Questions in Educational Domains*. https://doi.org/10.5281/ZENODO.6853085

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Arcas, B. A. y, … Natarajan, V. (2023). *Towards Expert-Level Medical Question Answering with Large Language Models* (arXiv:2305.09617). arXiv. https://doi.org/10.48550/arXiv.2305.09617

Sohail, A., & Zhang, L. (2024). Informing the treatment of social anxiety disorder with computational and neuroimaging data. *Psychoradiology*, *4*, kkae010. https://doi.org/10.1093/psyrad/kkae010

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 1–11. https://doi.org/10.1038/s41562-024-01882-z

Surameery, N. M. S., & Shakor, M. Y. (2023). Use Chat GPT to Solve Programming Bugs. *International Journal of Information Technology & Computer Engineering (IJITC) ISSN : 2455-5290*, *3*(01), Article 01. https://doi.org/10.55529/ijitc.31.17.22

Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General*, *153*(4), 1066–1075. https://doi.org/10.1037/xge0001547

Susnjak, T., Hwang, P., Reyes, N. H., Barczak, A. L. C., McIntosh, T. R., & Ranathunga, S. (2024). *Automating Research Synthesis with Domain-Specific Large Language Model Fine-Tuning* (arXiv:2404.08680). arXiv. https://doi.org/10.48550/arXiv.2404.08680

Tanay, B. A., Arinze, L., Joshi, S. S., Davis, K. A., & Davis, J. C. (2024). *An Exploratory Study on Upper-Level Computing Students' Use of Large Language Models as Tools in a Semester-Long Project* (arXiv:2403.18679). arXiv. https://doi.org/10.48550/arXiv.2403.18679

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Tian, H., Lu, W., Li, T. O., Tang, X., Cheung, S.-C., Klein, J., & Bissyandé, T. F. (2023). *Is ChatGPT the Ultimate Programming Assistant—How far is it?* (arXiv:2304.11938). arXiv. http://arxiv.org/abs/2304.11938

Tsai, H.-C., Huang, Y.-F., & Kuo, C.-W. (2024). *Comparative Analysis of Automatic Literature Review Using Mistral Large Language Model and Human Reviewers*. https://doi.org/10.21203/rs.3.rs-4022248/v1

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). *Do Large Language Models know what humans know?* (arXiv:2209.01515). arXiv. https://doi.org/10.48550/arXiv.2209.01515

Trott, S. (2024). Large Language Models and the Wisdom of Small Crowds. *Open Mind*, *8*, 723–738. https://doi.org/10.1162/opmi_a_00144

Turpin, M., Michael, J., Perez, E., & Bowman, S. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Advances in Neural Information Processing Systems*, *36*, 74952–74965.

Uittenhove, K., Martinelli, P., & Roquet, A. (2024). *Large Language Models in Psychology: Application in the Context of a Systematic Literature Review*. https://doi.org/10.31234/osf.io/nq4d2

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, *614*(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Vukojičić, M., & Krstić, J. (2023). ChatGPT in programming education: ChatGPT as a programming assistant. *InspirED Teachers' Voice*, *2023*(1), Article 1.

Vygotsky, L. S. (2012). *Thought and Language, revised and expanded edition*. MIT Press.

Wang, A., Morgenstern, J., & Dickerson, J. P. (2024). *Large language models cannot replace human participants because they cannot portray identity groups* (arXiv:2402.01908). arXiv. https://doi.org/10.48550/arXiv.2402.01908

Wang, J., Hu, H., Wang, Z., Yan, S., Sheng, Y., & He, D. (2024). Evaluating Large Language Models on Academic Literature Understanding and Review: An Empirical Study among Early-stage Scholars. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3613904.3641917

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., & Wen, Q. (2024). *Large Language Models for Education: A Survey and Outlook* (arXiv:2403.18105). arXiv. https://doi.org/10.48550/arXiv.2403.18105

Wang, X., Li, C., Chang, Y., Wang, J., & Wu, Y. (2024). *NegativePrompt: Leveraging Psychology for Large Language Models Enhancement via Negative Emotional Stimuli* (arXiv:2405.02814). arXiv. https://doi.org/10.48550/arXiv.2405.02814

Wang, Z. P., Bhandary, P., Wang, Y., & Moore, J. H. (2024). *Using GPT-4 to write a scientific review article: A pilot evaluation study* (p. 2024.04.13.589376). bioRxiv. https://doi.org/10.1101/2024.04.13.589376

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Xie, T., Wan, Y., Huang, W., Yin, Z., Liu, Y., Wang, S., Linghu, Q., Kit, C., Grazian, C., Zhang, W., Razzak, I., & Hoex, B. (2023). *DARWIN Series: Domain Specific Large Language Models for Natural Science* (arXiv:2308.13565). arXiv. https://doi.org/10.48550/arXiv.2308.13565

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review. *British Journal of Educational Technology*, bjet.13370. https://doi.org/10.1111/bjet.13370

Yax, N., Anlló, H., & Palminteri, S. (2024). Studying and improving reasoning in humans and machines. *Communications Psychology*, *2*(1), 1–16. https://doi.org/10.1038/s44271-024-00091-8

Ye, A., Maiti, A., Schmidt, M., & Pedersen, S. J. (2024). A Hybrid Semi-Automated Workflow for Systematic and Literature Review Processes with Large Language Model Analysis. *Future Internet*, *16*(5), Article 5. https://doi.org/10.3390/fi16050167

Zimmermann, R., Staab, M., Nasseri, M., & Brandtner, P. (2024). Leveraging Large Language Models for Literature Review Tasks—A Case Study Using ChatGPT. In T. Guarda, F. Portela, & J. M. Diaz-Nafria (Eds.), *Advanced Research in Technologies, Information, Innovation and Sustainability* (pp. 313–323). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-48858-0_25

Zhang, M., Press, O., Merrill, W., Liu, A., & Smith, N. A. (2023). *How Language Model Hallucinations Can Snowball* (arXiv:2305.13534). arXiv. https://doi.org/10.48550/arXiv.2305.13534

Zhao, S., Chen, S., Zhou, J., Li, C., Tang, T., Harris, S. J., Liu, Y., Wan, J., & Li, X. (2024). Potential to transform words to watts with large language models in battery research. *Cell Reports Physical Science*, *5*(3), 101844. https://doi.org/10.1016/j.xcrp.2024.101844

Zhou, K. Z., Kilhoffer, Z., Sanfilippo, M. R., Underwood, T., Gumusel, E., Wei, M., Choudhry, A., & Xiong, J. (2024). *'The teachers are confused as well': A Multiple-Stakeholder Ethics Discussion on Large Language Models in Computing Education* (arXiv:2401.12453). arXiv. https://doi.org/10.48550/arXiv.2401.12453