nature reviews psychology

https://doi.org/10.1038/s44159-023-00241-5

Perspective

Check for updates

Using large language models in psychology

Dorottya Demszky [1,10] ✉, Diyi Yang [2,10] ✉, David S. Yeager [3,4,10] ✉, Christopher J. Bryan [3,5], Margarett Clapper [3,4], Susannah Chandhok [6], Johannes C. Eichstaedt [7,8], Cameron Hecht [3,4], Jeremy Jamieson [9], Meghann Johnson [3], Michaela Jones [3], Danielle Krettek-Cobb [6], Leslie Lai [6], Nirel JonesMitchell [3], Desmond C. Ong [3,4], Carol S. Dweck [7], James J. Gross [7] & James W. Pennebaker [4]

Abstract

Large language models (LLMs), such as OpenAI's GPT-4, Google's Bard or Meta's LLaMa, have created unprecedented opportunities for analysing and generating language data on a massive scale. Because language data have a central role in all areas of psychology, this new technology has the potential to transform the field. In this Perspective, we review the foundations of LLMs. We then explain how the way that LLMs are constructed enables them to effectively generate human-like linguistic output without the ability to think or feel like a human. We argue that although LLMs have the potential to advance psychological measurement, experimentation and practice, they are not yet ready for many of the most transformative psychological applications – but further research and development may enable such use. Next, we examine four major concerns about the application of LLMs to psychology, and how each might be overcome. Finally, we conclude with recommendations for investments that could help to address these concerns: field-initiated 'keystone' datasets; increased standardization of performance benchmarks; and shared computing and analysis infrastructure to ensure that the future of LLM-powered research is equitable.

Sections

Introduction

Conceptual and methodological foundations

Concerns about applying LLMs to psychology

Outlook

---

[1]Graduate School of Education, Stanford University, Stanford, CA, USA. [2]Department of Computer Science, Stanford University, Stanford, CA, USA. [3]Texas Behavioral Science and Policy Institute, University of Texas at Austin, Austin, TX, USA. [4]Department of Psychology, University of Texas at Austin, Austin, TX, USA. [5]Department of Business, Government, and Society, University of Texas at Austin, Austin, TX, USA. [6]Google, LLC, Mountain View, CA, USA. [7]Department of Psychology, Stanford University, Stanford, CA, USA. [8]Institute for Human-Centered AI, Stanford University, Stanford, CA, USA. [9]Department of Psychology, University of Rochester, Rochester, NY, USA. [10]These authors contributed equally: Dorottya Demszky, Diyi Yang, David S. Yeager. ✉e-mail: ddemszky@stanford.edu; diyiy@stanford.edu; yeagerds@austin.utexas.edu

Introduction

Language is a core aspect of all areas of psychological science. For instance, language is a primary means through which children form mental models of the world (developmental psychology). These mental models are revealed through verbal explanations of events (cognitive psychology) and interactions with others (social psychology). Language can reveal disordered mental models that could be targeted by treatments (clinical psychology) or communicate emotions and guide emotion regulation strategies (affective psychology), which have important implications for physical and mental well-being (health psychology). Individual differences in traits such as agreeableness, extraversion or narcissism are expressed in spoken and written communication styles (personality psychology). Finally, research is conveyed through text in scholarly papers, which then serve as objects of study in systematic reviews (metascience).

Technological advances in natural language processing in the past two decades have made it easier than ever for psychologists to incorporate language analysis in their work. As computational power and the availability of data have grown, language analysis has progressed from humans counting words in text, starting in the 1960s[1,2], to latent semantic analysis in the 1990s[3,4] and to automated algorithms for analysis of text on the internet[5–8] in the 2010s. For instance, one popular algorithm (the Linguistic Inquiry and Word Count program[9]) enables researchers to examine the frequency of occurrence of different categories of words and assess how they predict outcomes such as physical and psychological health. Another prominent example is topic modelling, which uncovers themes in documents by analysing word co-occurrence patterns (via Latent Dirichlet allocation[10]). The advent of LLMs, such as OpenAI's generative pre-trained transformer (GPT) models[11,12], Google's LaMDA and Bard AI[13] (which can both generate and score text in human-like ways), will probably push the field in new directions.

LLMs are massive statistical models that predict the next word, phrase, sentence or paragraph, given an input. They are novel and powerful for two key reasons. First, LLMs go beyond popular word-counting bag-of-words methods (such as Linguistic Inquiry and Word Count) by approximating meaning in context, rather than assuming that words have a static meaning across contexts. This feature of LLMs is important because intended meaning does not have a one-to-one mapping with words. The context of words (topic, register, style, sarcasm or metaphor) influences their meaning[14,15]. Thus, accounting for meaning in context can more authentically represent human language[4]. Second, LLMs are trained on an unprecedented amount and variety of text, including hundreds of billions of utterances from online digital sources, such as web pages, books, magazines, social media, Wikipedia and online forums[11,12]. Thanks to better modelling and data, LLMs can measure subtle, context-dependent expressions of mental states, generate text that is human-like, synthesize information from a wide range of sources, engage in natural dialogue and simulate different linguistic styles and personas.

To illustrate these capabilities, Fig. 1 shows how an off-the-shelf LLM performed various tasks related to a nervous high school student applying to college. This college-advising example points to potentially broad applications of LLMs across diverse areas of psychology. LLMs can be used to query people's mental models of themselves and their environments (social and cognitive psychology), infer individual differences in coping styles (personality psychology), or help people to reappraise stressful experiences (affective and clinical psychology). Moreover, the text in Fig. 1 was generated in seconds, showing that LLMs can accomplish tasks relevant to psychologists nearly instantaneously. Thus, LLMs unlock possibilities for scale and efficiency in psychological research and practice that were unthinkable just a few years ago.

Extrapolating from examples such as this, researchers might envision applying LLMs to measurement (for example, to probe psychological patterns in massive bodies of text), experimentation (to rapidly generate stimuli and responses to those stimuli to improve the granularity of factorial designs) and practice (to simulate realistic conversations to train human therapists, counsellors, teachers or managers). Table 1 previews use cases across subfields in each of these three categories of applications. We generated these use cases by consulting experts in several areas of psychology about how LLMs could make their work more efficient or wholly different. These use cases are not exhaustive, but rather illustrate the breadth of what might be achieved if the field is successful at adapting LLMs for psychological purposes.

The potential uses of LLMs are as tantalizing for the lay public as they are for researchers. For example, the public rushed to use LLMs such as ChatGPT for mental health screening and therapy[16], and a nationally representative survey found that 40% of US teachers were using ChatGPT weekly for lesson planning just three months after its public release[17]. The question is not whether LLMs will influence everyday life and psychology. Instead, the question is whether – and on what timescale – the field will apply a nuanced understanding of LLM technology to developing and disseminating tools so that psychologists can harness its power for their research, while avoiding the harm that might come from premature applications.

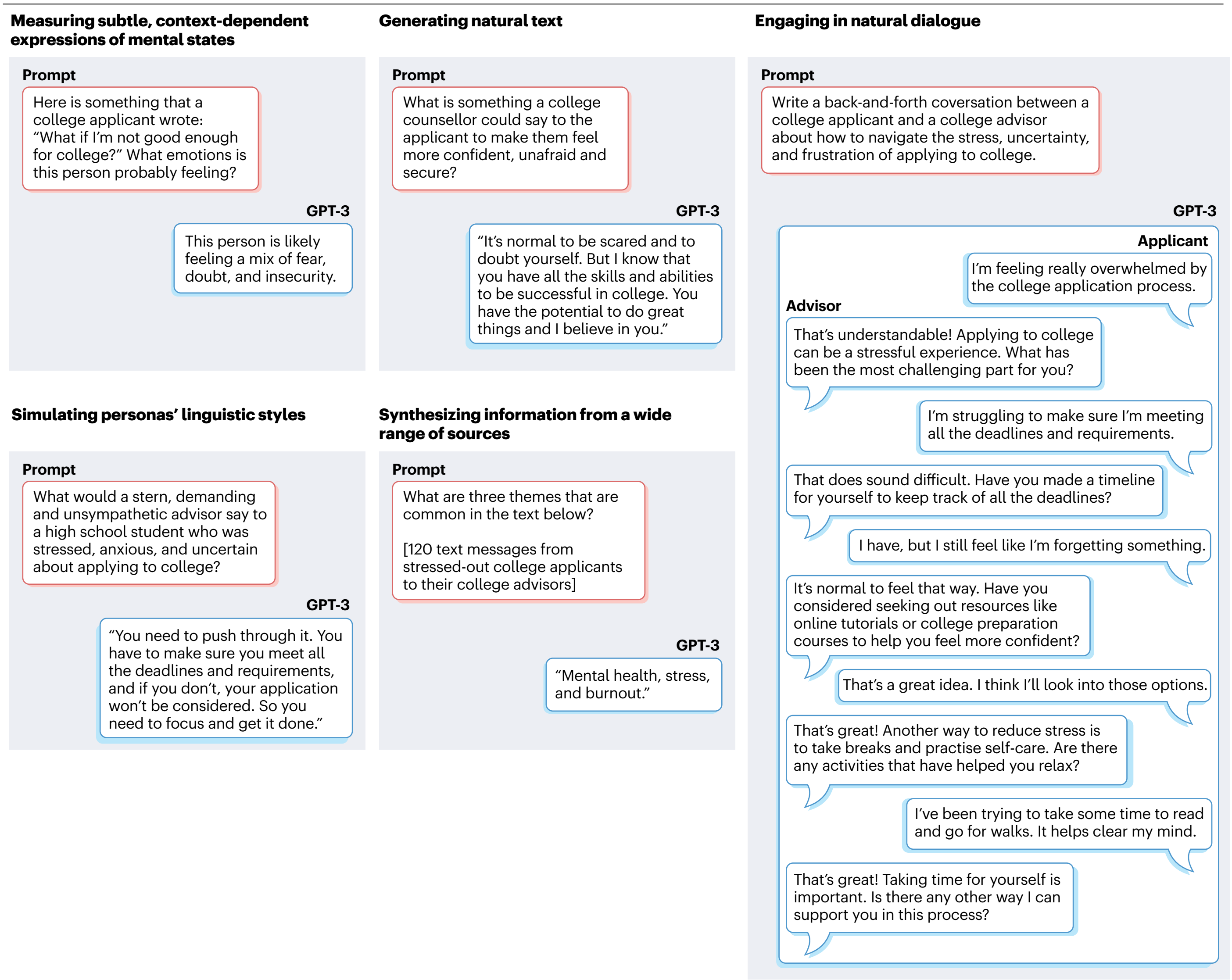
In this Perspective, we explore issues related to using LLMs in psychology. First, we describe how LLMs work, and what this information means for their applications in psychology. Next, we address concerns about using LLMs for psychological purposes, along with potential solutions to these concerns. It is important to raise these concerns because if they are left unaddressed LLMs might be used in ways that cause harm to the public. Members of underrepresented groups in particular could suffer harm owing to biases in how AI models are trained and implemented. We conclude by discussing how collaborative, team-science initiatives could advance psychological science and inform an ethical future for the applications in practice of LLMs.

Conceptual and methodological foundations

As with many new technologies, LLMs have the potential to change psychology research, but also to be misunderstood. Thus, it is critical to understand what LLMs are and what they are not. Here we provide a brief overview of how LLMs are trained, how they work and how they can be adapted to the use cases presented in Table 1. We summarize currently available software tools for using LLMs in Box 1.

How the models are fitted

LLMs are designed to reproduce word co-occurrence patterns that were found in their training data. They have become very good at this task owing to immense quantities of training data and complex architecture. The most common model architecture for LLMs is the 'transformer', which forms the backbone of modern models such as BERT, GPT, Bard and LaMDA. These LLMs are essentially massive systems of nonlinear regression equations (specifically, neural network machine learning models). These models often have millions of parameters that were calculated by taking sentences as predictors (an X in a regression equation) and masked-out words or the next sentence as an outcome (Y). The prediction error from one fit is then used to update the models' parameters (via backpropagation), and the process is repeated



Measuring subtle, context-dependent expressions of mental states

Prompt

Here is something that a college applicant wrote: "What if I'm not good enough for college?" What emotions is this person probably feeling?

GPT-3

This person is likely feeling a mix of fear, doubt, and insecurity.

Generating natural text

Prompt

What is something a college counsellor could say to the applicant to make them feel more confident, unafraid and secure?

GPT-3

"It's normal to be scared and to doubt yourself. But I know that you have all the skills and abilities to be successful in college. You have the potential to do great things and I believe in you."

Simulating personas' linguistic styles

Prompt

What would a stern, demanding and unsympathetic advisor say to a high school student who was stressed, anxious, and uncertain about applying to college?

GPT-3

"You need to push through it. You have to make sure you meet all the deadlines and requirements, and if you don't, your application won't be considered. So you need to focus and get it done."

Synthesizing information from a wide range of sources

Prompt

What are three themes that are common in the text below?

[120 text messages from stressed-out college applicants to their college advisors]

GPT-3

"Mental health, stress, and burnout."

Engaging in natural dialogue

Prompt

Write a back-and-forth coversation between a college applicant and a college advisor about how to navigate the stress, uncertainty, and frustration of applying to college.

GPT-3

Applicant

I'm feeling really overwhelmed by the college application process.

Advisor

That's understandable! Applying to college can be a stressful experience. What has been the most challenging part for you?

I'm struggling to make sure I'm meeting all the deadlines and requirements.

That does sound difficult. Have you made a timeline for yourself to keep track of all the deadlines?

I have, but I still feel like I'm forgetting something.

It's normal to feel that way. Have you considered seeking out resources like online tutorials or college preparation courses to help you feel more confident?

That's a great idea. I think I'll look into those options.

That's great! Another way to reduce stress is to take breaks and practise self-care. Are there any activities that have helped you relax?

I've been trying to take some time to read and go for walks. It helps clear my mind.

That's great! Taking time for yourself is important. Is there any other way I can support you in this process?

Fig. 1 | Examples of LLM functionality. Prompts and GPT output for various tasks related to a nervous high school student applying to college. GPT output was generated in March 2023 using text-davinci-003 in the OpenAI playground with default settings (temperature 0.7, length 256).

until the parameters sufficiently minimize the difference between the model output and the target text. This process is broadly referred to as pre-training (Fig. 2a). The newest models such as ChatGPT and GPT-4 involve additional pre-training steps, including instruction tuning and reinforcement learning from human feedback to further improve the model's ability to generate coherent and relevant language[18,19].

Off-the-shelf LLMs are powerful, but they cannot fulfil their potential for psychology research without further adaptation. For example, in the LLM-generated dialogue on the right of Fig. 1, the LLM-generated college advisor asked whether the LLM-generated applicant had created a 'timeline' or taken 'online courses'. These are reasonable things for an advisor to say but are not necessarily good advice. In fact, according to a large, rigorous, randomized experiment funded by the US Department of Education, this kind of college-advising guidance has no effect on college persistence compared to a control condition[20]. In summary, the LLM generated on-topic, grammatically correct useless information.

Furthermore, the LLM-generated advisor encourages the applicant to 'reduce stress', 'take breaks' and practice 'self-care'. These are common themes in advice from advisors. However they are generally not what research on stress shows is helpful or what might be most appropriate given the context[21,22]. The LLM seems to have acquired a stress-is-debilitating mindset[23] — the notion that stress always harms performance and health. This mindset is contradicted by research that emphasizes the importance of understanding the potentially enhancing role of stress. That is, the body's stress response keeps people alive by helping to mobilize energy for performance (for example, getting more oxygenated blood to brains and muscles)[21,22].

Thus, the college-advising example in Fig. 1 illustrates that LLMs are not designed to predict what language should come next, according

Table 1 | Example use cases for LLMs in different areas of psychology

Use case	Area	Example
Measurement	Cross-area	Measure whether, and to what extent (for example, on a scale of 1 to 10), a text (such as an email or text message) contains high-inference, context-dependent constructs[30,90], such as fear or anxiety, belonging, implicit bias[91,92] or suicide risk[93]. Probe the LLM to gain insight into how concepts and ideas hang together in the dominant culture that produced the training data — that is, conduct a kind of 'digital ethnography' — by prompting the LLM with standardized questions or vignette-based tasks[25], or by measuring the associations among concepts[94].
	Affective	Score text on massive online platforms (such as Reddit) written by users before, during and after a traumatic period (for example, a break-up, death or war). By tracking text on social media, it is possible to measure the social and emotional costs of large-scale health upheavals in almost real time[95,96]. Doing so could help affective scientists identify the thinking patterns, social connection, self-talk and appraisals associated with coping in the real world[97].
	Cognitive	Identify people's mental models for telling or hearing a story or a causal explanation, which could reveal basic cognitive processes underlying how people process narrative or explanatory information[98–100]. Probe an LLM to generate concepts associated with different essentialist beliefs about categories to understand the origins and implications of essentialist reasoning in natural dialogue[101].
	Developmental	Estimate the difficulty of learning different words based on LLMs' prediction probabilities[102]. Analyse narrative retellings produced by children at different developmental stages and with different conditions (such as autism) for accuracy, intelligibility, grammatical complexity and inclusion of critical events to better understand their narrative comprehension and production skills[103].
	Metascience	Facilitate scientific reviews and meta-analyses[104] by assessing the meaning in context of scientific keywords in abstracts, which could reduce excessive 'hits' during the literature search process. Code manuscript features (for example, study quality factors)[105] for use in meta-regression analyses. For example, LLMs could help to discover red flags for *P*-hacking[75] in a contextually attuned way, rather than using keyword searches (such as 'covariate' or 'transformation').
Experimentation	Cross-area	Generate experimental stimuli that show different gradients of a construct of interest (such as authenticity or growth mindset) in a variety of social groups (such as among managers, police, parents and teachers) to use in a large multi-factorial vignette experiment that could identify the active ingredients for a psychological manipulation. Generate hundreds of alternative items for self-report instruments using a few examples as prompts, then use established methods (such as cognitive pre-testing, factor analysis, and convergent and predictive validation) to evaluate the properties of the new items, and to develop an optimal final instrument[76]. Generate synthetic datasets of LLM-generated responses in humanless pilot experiments[29,32,106]. Researchers could conduct statistical analyses on synthetic responses to adapt treatments, estimate effect sizes and conduct power analyses.
	Social	Score language on social media or other platforms for potentially harmful language (for example, prejudice, misinformation or extremism[107,108]), and use the resulting algorithms to evaluate the real-world impact of interventions to reduce harmful language on social media.
Measurement and experimentation	Health	Rapidly generate individually tailored 'nudges' to promote healthy habits, matching each participant's demographics and values reported on a baseline survey[109], to causally test the sources of heterogeneous treatment effects. Score text written by users before, during and after a challenging physical health period, and prompt them to reconsider their health-related thoughts or behaviours.
	Personality	Score an email or text message corpus to identify what distinguishes the way that people high on a personality trait (for example, extraversion) talk about themselves compared to people low on that trait. Then train an LLM to simulate conversation partners who are high versus low on the trait to causally test the impact of a personality trait on interaction quality.
Practice	Clinical	Build a simulated environment for mental health providers-in-training to practise helping patients to reappraise problematic construals[65,110]. Use a fine-tuned LLM to generate example responses to patients in distress (that the therapist can edit) which could help therapists (or even peers[66]) to deliver care more efficiently and at scale.
	Education	Provide automated feedback to teachers on high-leverage, student-centred teaching practices (such as growth mindset, belonging or relevance) to improve instruction and thereby improve student motivation and achievement[65,67,111,112].

Each of these suggested applications requires testing and validation to ensure representation, fairness and robustness.

to the truth of the matter, let alone according to some standard from psychological theory or evidence. They are not generating advice with the listener's well-being — or the scientific literature — in mind. They are simply parroting what the training data tended to say about the concept, so that the dialogue sounds natural. In other words, if the cultural context that produced the training text tends to give bad advice about applying to college, then the LLM will probably parrot common but ultimately unhelpful advice[24].

That said, the biases in LLMs' output might be interesting to research. Researchers might ask: what does it say about human culture that the LLM learned a stress-is-debilitating mindset from its training data? Because LLMs are trained on massive bodies of human language, probing LLMs provides a window into how concepts and ideas hang together in the dominant culture that produced the training data[25]. This feature could enable researchers to trace the collective knowledge of the people who generated the training data.

Adapting LLMs for specific purposes

Although off-the-shelf LLMs might provide insight into the way people on the internet see concepts such as a stress-can-be-enhancing mindset, scientists typically want to study constructs using scientific conceptions of those constructs. Thus, using LLMs to advance scientific theory and to fashion state-of-the-art interventions requires a different approach than relying on off-the-shelf LLMs. In particular, psychologists will need to additionally train (or 'tune') LLMs on high-quality data curated by experts for a given construct. Moreover, they will need to conduct evaluation experiments to see whether the LLM-generated data are as effective as intended. This need for additional training and evaluation is important to emphasize because much of the public's conversation about LLMs has focused on off-the-shelf performance. However, the LLMs that will be most effective are likely to require additional training with datasets specific to a particular domain. For example, DocsGPT is an LLM that is pre-trained on medical prose to help with clinical correspondence and charting[26]. Here we describe how fine-tuning and prompt-tuning can be used to tune

Box 1

Tools to use LLMs

Here we describe tools that can be used to harness LLMs with minimal expertise. However, a limitation of using tools that require little understanding is that incorrect defaults might be used. In addition, the field is changing rapidly. The best tools might be outdated by the time this paper is published.

APIs and online interfaces

A growing number of application programming interfaces (APIs) and online interfaces enable users to interact with pre-trained LLMs. The main benefit of these APIs and interfaces is that they tend to be accessible to users with no technical background. For example, OpenAI provides a simple textbox for GPT-3 and ChatGPT in which the user can interact with the LLM. In addition, the user does not need to worry about the computational needs required for hosting and serving large models.

The main disadvantages of using APIs and online interfaces are limited customization, transparency, and in many instances, there is a financial cost. Online platforms tend not to allow for fine-tuning (only prompt-tuning) and the input format is currently limited to text written in a box (for example, users cannot upload a file with multiple variables). Although some APIs do allow fine-tuning, they still do not allow users to inspect the models' parameters, which can be important for interpretation. In addition, most APIs and online interfaces are behind a paywall.

Downloading and hosting models

We recommend downloading or hosting a model oneself when smaller LLMs are sufficient and when data security and control over the models is a high priority. Having more control means more room for customization, evaluation of 'black box' model parameters, greater data privacy and, in some cases, decreased costs. Technical expertise might become less of a barrier to working with downloaded models owing to the development of easy-to-use packages and libraries (see below). However, the most powerful models are not downloadable or are only accessible via APIs such as OpenAI's ChatGPT.

Hosting large models can incur large costs and hardware requirements. LLMs usually need access to graphical processing units (GPUs) or tensor processing units during model training and application. These types of processors are particularly well suited to the underlying computations of transformer-based models. They require fairly large amounts of memory to operate. Smaller LLMs (BERT, RoBERTa, GPT-2), which perform well for many measurement tasks, can be implemented with fewer resources, some only with a regular laptop. There are also free platforms, such Google Colab (a cloud-hosted Jupyter notebook for Python code) that provide access to basic GPUs that can be used to work with pre-trained LLMs.

Useful software packages and libraries

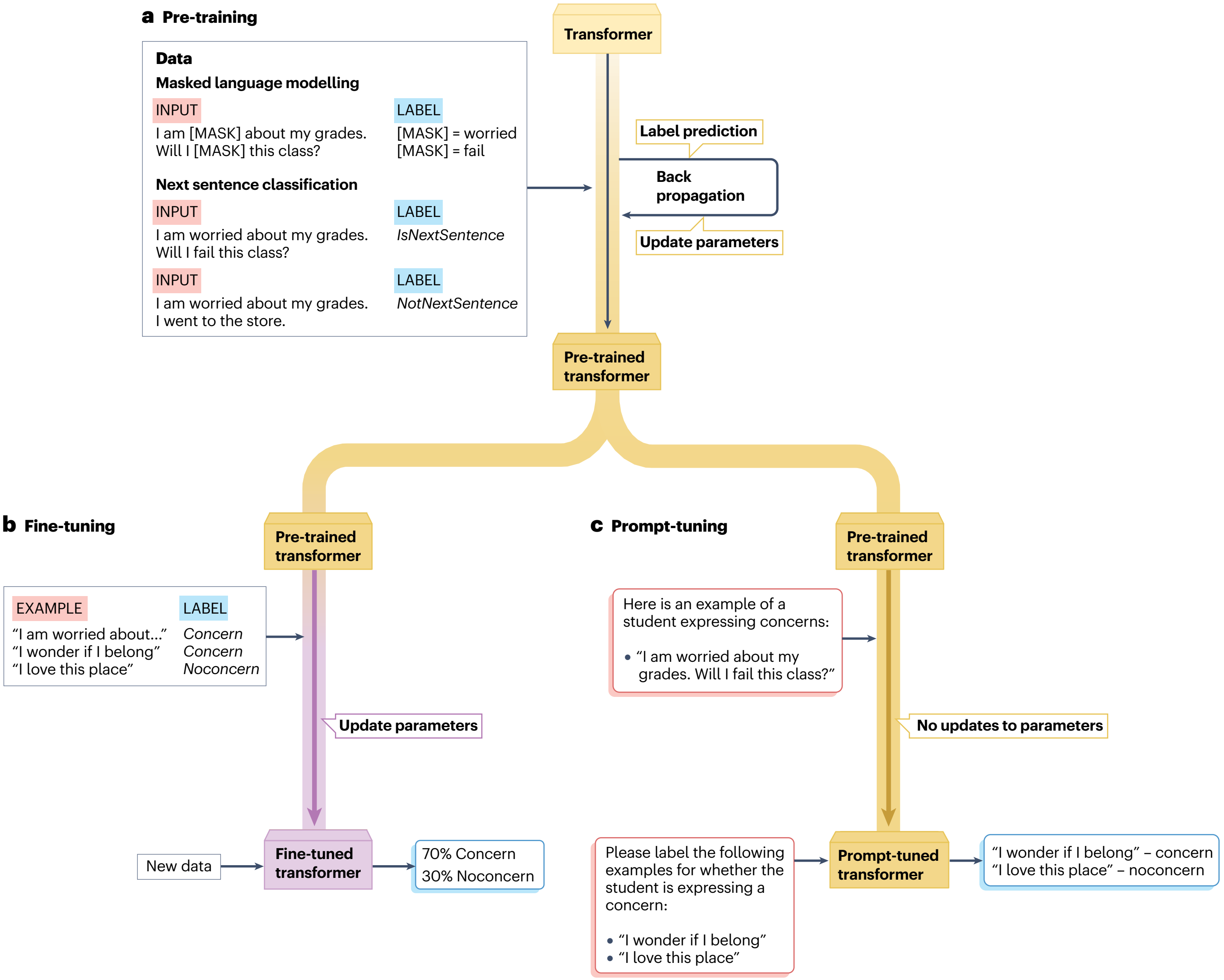
Fully pre-trained, off-the-shelf LLMs can be used in a browser or accessed through an R package (for example, text). Deeper engagement with language models might require work with Python libraries, such as Hugging Face. These are becoming easier to use every day. A good trade-off between complexity and the range of possible applications is provided by the open-source Differential Analysis ToolKit (DLATK). This is a Linux-based library that provides a user-friendly interface to many natural language processing Python libraries, including those that provide access to LLMs[113]. LLMs have already become integrated in products such as Google Sheets (for example, SheetGPT.ai), and we anticipate that LLMs will become increasingly more accessible to the consumer through intuitive interfaces.

Example Python packages include:

- openai (a Python interface over OpenAI's API endpoints)
- DLATK (a human text analysis package written in Python that supports transformer-based models from Hugging Face, in addition to other models)
- langchain (a Python framework to develop LLM-powered applications, which supports various LLMs, including open-source models hosted in the Hugging Face model hub)
- chainlit (a Python framework that uses langchain to build ChatGPT-like LLM apps).

Example R packages include:

- TheOpenAIR (an R package to integrate OpenAI's GPT models into R workflows)
- askgpt (another R package built around OpenAI APIs to access the GPT models in R)
- reticulate (a Python interface written in R that allows us to import Python modules and call functions, which can be used to work with LLaMA and other LLMs, and langchain)
- gptstudio and gpttools (RStudio add-ins that incorporate LLMs into R programming)
- chatgpt (an R coding assistant built around OpenAI APIs).



INPUT	LABEL
I am [MASK] about my grades. Will I [MASK] this class?	[MASK] = worried [MASK] = fail
I am worried about my grades. Will I fail this class?	*IsNextSentence*
I am worried about my grades. I went to the store.	*NotNextSentence*

EXAMPLE	LABEL
"I am worried about..."	*Concern*
"I wonder if I belong"	*Concern*
"I love this place"	*Noconcern*

Here is an example of a student expressing concerns:

- "I am worried about my grades. Will I fail this class?"

Please label the following examples for whether the student is expressing a concern:

- "I wonder if I belong"
- "I love this place"

"I wonder if I belong" – concern
"I love this place" – noconcern

Fig. 2 | Pre-training, fine-tuning and prompt-tuning of LLMs. a, During pre-training, a dataset of sentences is used as input to the LLM (marked 'transformer' to represent the most common architecture of current LLMs). The inputs (for example, masked-out words, paired sentences) are processed automatically and the LLM is optimized to reconstruct the original text (for example, to predict the masked word or the next sentence). **b**, Fine-tuning is the subsequent process of refining the model on a smaller, more specific dataset to adapt it to a particular domain or task. In this example, an LLM is fine-tuned on a dataset of sentences labelled as expressing a concern or not (noconcern) by adding an additional classifier layer on top of the pre-trained model (an x-by-2 matrix, where x is the size of vector output by the transformer and 2 is the number of labels (concern, noconcern)). Then, the fine-tuned model can be applied to a new dataset to predict what percentage of sentences in that dataset reflect a concern. **c**, Prompt-tuning involves optimizing the input prompts to guide the pre-trained model's behaviour on specific tasks without changing the model's parameters. Whereas pre-training forms the base for general language understanding, fine-tuning and prompt-tuning adapt the model to a specific task with and without changing the model's parameters, respectively.

LLMs to specific domains (for example, parent–child interactions) or constructs (for example, optimal stress coping or a growth mindset) that are not represented scientifically in the public's language or in the LLM's training data.

Fine-tuning. Fine-tuning (Fig. 2b) involves giving the pre-trained LLM new and more specific training data to adjust the model's parameters for a specific task. Fine-tuning data can come in many forms. Generally, the larger and the more representative the data is of the target task and domain, the better the fine-tuned model will perform. The new data might be a set of texts annotated by experts for a 'stress-is-debilitating' versus a 'stress-can-be-enhancing' mindset, high versus low narcissism, metaphors that represent causal relationships, or the expected responses to a given query. Annotations do not always need to come from experts. They can include lay annotations (which might have bias) or any kind of metadata, such as reactions to social media posts or the demographic data of those who produced the text. Unlabelled text can also be used to improve the models' performance in a particular domain that is under-represented in the training data, such as language from parent–child interactions or language from people with a speech-related disability.

Fine-tuning LLMs is powerful because it requires orders-of-magnitude less data than would be needed to train a specialized model from scratch (as few as 5–100 annotated examples may be needed to fine-tune versus potentially hundreds of thousands of examples to train a new model). Fine-tuning a pre-trained LLM therefore substantially reduces the amount of hand-annotated data needed (Box 2).

Fine-tuning also has its drawbacks. First, fine-tuning is still computationally expensive. Powerful servers are required to update the LLM's parameters. Second, high-quality annotated data to use in fine-tuning can be difficult to come by. For example, coders might not agree on what counts as a narcissistic statement, a growth mindset statement, or bad advice for coping with stress. Finally, the fine-tuned model might retain biases that were inherent in the LLM's foundational training corpus. It could even become more biased depending on who did the fine-tuning (for example, researchers without expertise or who fine-tune from their own potentially biased viewpoints). As discussed below, a critical task for the field is to curate large, reliable annotated datasets of key psychological constructs while minimizing unwanted biases.

Prompt-tuning. A second approach to customizing an LLM for a specific purpose is prompt-tuning (also called in-context learning)[27,28] (Fig. 2c). Prompt-tuning is becoming popular because the LLM does not need to be re-fitted (that is, the underlying model parameters are not changed). Instead, the type of output the LLM generates is tailored to match user-generated prompts. Thus, the customization can happen quickly. For example, a user interested in fostering a stress-can-be-enhancing mindset could prompt-tune an LLM with the following prompt: "Here is some good advice to give to people who are very stressed. [user enters one or two examples written by the user or an expert]. What is a way to give this advice to someone who is stressed about work?" The LLM would then output one or two examples. The user could then ask (in the same session with the LLM): "What is a way to give this advice to someone who is stressed about relationships?" The LLM would then output one or two examples. Thus, even just a couple of user-generated examples can help the model to extrapolate structural and semantic features (Box 3) and use them to generate a new set of examples in response to similar user queries[29].

Prompt-tuning faces the same limitations with respect to high-quality annotated data as fine-tuning, but fewer examples (for example, one to ten) are needed. The key limitation of prompt-tuning is that it might not be very effective at addressing the many subtleties and complexities of a specific psychological construct, unlike a fine-tuned model whose parameters are adjusted based on a high-quality annotated dataset. Prompt-tuning is also less likely than fine-tuning to overcome biases in the data because its parameters are not updated. Finally, a challenge with prompt-tuning is that the LLMs' outputs will be sensitive to how the prompt was worded. Thus, prompt engineering (in which users vary the prompts given to the LLM to achieve a better output for a given purpose) is a key aspect of prompt-tuning that requires subtle expertise[30,31].

LLM providers such as OpenAI, Meta or Google prioritize model customization via prompt-tuning rather than fine-tuning; their latest models (OpenAI's GPT-4 and Google's Bard AI) cannot be fine-tuned by public users. One potential reason for preventing users from fine-tuning is that fine-tuning might interfere with the provider's control of the model, including guardrails that prevent harmful outputs such as hate speech. Anticipated enhancements might elevate the effectiveness of prompt-tuning to that of fine-tuning for certain tasks. For example, OpenAI allows for increasingly detailed instructions via system messages (such as step-by-step directions for responding to an adolescent who is anxious about applying to college) that can guide the model's behaviour throughout a conversation and reduce issues associated with earlier prompt-tuning approaches (such as forgetting instructions). However, using open-source models that can be fine-tuned (such as Meta's LLaMA2) is recommended for researchers who require more thorough customization.

Concerns about applying LLMs to psychology

The widespread enthusiasm for LLMs should be tempered by an awareness that they are not actually simulating human intelligence[29,32,33]. LLMs simply predict the next phrase or sentence, given what they have been

Box 2

Fine-tuning for measurement

Demszky and colleagues (D.D., unpublished work) fine-tuned a language model (RoBERTa)[114] to identify language that reflected first-year college students' worries and attitudes about belonging in their new school environment. The researchers annotated a subset of 650 short, first-person essays written by participants in a large experiment testing the effect of a social belonging intervention[70] at over 26 colleges and universities[115].

The essays were coded by trained graduate student annotators for themes related to belonging, specifically whether the essays portrayed belonging concerns as common; whether they explored changes in belonging over time; and whether students connected these ideas to their personal lives. Next, the subsample of 650 essays were used to fine-tune an LLM so that it was 'biased' towards these three language features. Finally, the entire set of approximately 11,000 essays were scored by the LLM for their tendency to endorse these belonging themes.

Analysis of the LLM-scored essays found that students whose essays reflected more themes of the belonging intervention (reflecting on belonging concerns as being common and temporary) did better in college over the year. Consistent with theories of identity threat, this effect was greater among students from minoritized groups[70,115]. Some next steps for this research might be to unpack the 'black box' of what the LLM was using for its scoring.

Box 3

Prompt-tuning for experimentation

Clapper and colleagues[116] used prompt-tuning to evaluate the ability of two LLMs (GPT-3 and LaMDA) to generate growth mindset language (language supporting students' belief that intellectual ability can be developed[117]) in the form of an Algebra I teacher's speech describing how to do well in the class that year[118]. The study had three conditions: speeches generated by un-tuned LLMs; speeches written by expert growth mindset math teachers; and speeches written by LLMs that were prompt-tuned with the expert human speeches from the second condition. These speeches were then evaluated by a new sample of about a thousand adolescents. In the two LLM conditions, the prompt was: "Write a speech for the first day of Algebra class that tells your students what learning math is like."

A typical speech generated by the un-tuned LLM was: "Hello students! Welcome to Algebra! For those of you who are new to this class, Algebra is all about solving equations. We will be using algebraic methods to solve equations and to find out information that is not given to us in the equation. Algebra is a challenging class, but it is also a lot of fun. I promise that if you put in the work, you will be successful.". Although this speech is positive and encourages effort, it was not rated as being high in growth mindset by the adolescent evaluators because it did not address students' fixed mindset fears (that is, that when they struggle it means they are 'bad at maths').

A typical speech generated by the expert-prompt-tuned LLM was: "Welcome to my Algebra 1 class! I am eager to start learning with all of you. For those of you who may be feeling a little intimidated by math, I want to let you know that you are capable of learning this material. It may be difficult at first, but with effort and determination, you will be able to understand it. By the end of the year, you will be able to solve problems that once seemed impossible. So let's get started on this journey together and see what we can achieve!" This speech was rated higher than the un-tuned LLM speech on growth mindset by adolescents because it explained the path towards improvement.

One of the most striking findings was that adolescents' ratings of the prompt-tuned examples were not significantly different ($P > 0.05$) from ratings of the human expert examples — even though the former were generated in seconds, whereas the latter were generated over a three-day period. Furthermore, the expert-prompt-tuned speeches made up 75% of the gap between un-tuned LLM and human expert speech ratings. These findings provide direct evidence that LLMs can generate novel text that is similar to a precise psychological construct with only five training examples. This study also suggests a promising approach to using LLMs for psychological research when the base model gets the construct wrong: use expert judgements to select exemplary text, and then prompt-tune the LLM to approximate that expert corpus.

exposed to in the training data. Consequently, they tend to output an 'average' of what the internet or popular books tend to say (with some constraints, for example LLM providers have put ethical 'guardrails' in place to prevent hate speech). This is an advantage for research areas that seek to leverage and study human cognition and culture as represented in the training data. With appropriate prompting and tuning, LLMs can also become a fascinating sandbox for conducting metascientific analyses, creating interactive interventions and generating a wide range of stimuli. At the same time, LLMs have limitations with respect to all these applications, because they do not possess the non-language-specific cognitive capacities required for modelling thought[34]. Psychological studies purporting to evaluate the psychological reasoning 'abilities' of LLMs (for example, heuristics and biases or theory of mind) could be subject to serious (if not fatal) validity concerns[35,36]. In this section we describe four areas of concern that currently limit the use of LLMs by psychologists.

How to evaluate LLMs

The public launch of ChatGPT in November of 2022 led to a flood of lay evaluations of LLMs. Many involved asking the LLM leading questions to goad it into saying something problematic[37,38]. The issues in the model's outputs led many observers to conclude that LLMs are not very 'good'. However, a few months later the developers of the GPT models released a report showing that the latest LLM could pass the Bar Exam (which certifies US lawyers), ace the Graduate Record Examination (GRE; used in applications to many US graduate schools), and earn a 5 (the highest score) on every Advanced Placement test[39]. These reports would seem to suggest that LLMs are actually quite good at tasks that require human-like intelligence.

Importantly, the standards for what constitutes 'good' LLM performance to psychologists interested in using LLMs to perform scientific or professional psychology tasks are different from the standards encoded in the transformer models that trained the LLM (namely, that the next set of words is plausible). For the computer scientists who developed the LLMs, an LLM might be 'good' if it tends to output text that is grammatical, coherent and natural-sounding, given the input. Optimizing for these criteria can account for both the harmful speech the LLM can generate (because humans often use harmful speech) and its high test performance. Both metrics come from the same underlying abilities. But neither of these metrics capture the criteria needed for psychologists to trust LLMs. Said differently, computer scientists have tended to evaluate the functionality of features, but psychologists usually want to evaluate the effects of those features on human thought and behaviour. Here we introduce two high-level approaches to evaluating LLMs: expert evaluation and impact evaluation.

Expert evaluation. A standard way to evaluate LLMs is to measure the extent to which their output score or text agrees with those from human domain experts. Here, 'experts' are broadly construed as those who are representative of or familiar with the target construct one is trying to measure or generate. When using LLMs for psychological measurement, trained research assistants and LLMs could score the same texts for particular constructs, and then compute agreement between their scores (Box 2). When using LLMs for text generation, experts and the LLM could be given the same prompt, and their responses could be compared using content analysis (Box 3).

Expert evaluation has long been used by computer scientists to refine LLMs' capabilities for generating human-like text. This approach differs in important ways from evaluating the LLMs' abilities to generate psychologically helpful language. In general, people tend to agree on what counts as human-like language and demonstrate moderate to high agreement on simple codes such as coherence, grammaticality and relevance[40,41]. But people do not always agree on whether a message is psychologically helpful[42] because of the intersubjectivity problem[43], which arises from the basic social psychological principle that a person cannot fully know how another person is construing the meaning of a situation. This intersubjectivity problem creates three interrelated issues for expert evaluation of LLMs.

First, coders are third parties making guesses about how someone else would feel if they received the text generated by the LLM. These guesses can be wrong when the coders are not attuned to the context. For example, consider the use case shown in Fig. 1 of identifying text that best comforts a stressed-out college applicant. Third-party coders might not be aware of how broader concerns (for example, about paying for college or disappointing one's parents) are influencing the deeper meaning that a vulnerable applicant is deriving from the advisor's text.

Second, the intersubjectivity problem is compounded by cross-group judgements (for example, across race or ethnicity, gender or political ideology) when raters are asked to draw inferences on behalf of people with very different identities, experiences or beliefs from their own. This problem is further exacerbated across power differences. Higher-power groups are generally less attuned to the subjective perspectives of lower-power minority groups[44]. Thus, third-party coders from majority groups could be especially inaccurate at guessing how individuals from minority groups would interpret LLM-generated text.

Third, when coding for nuanced psychological constructs, sometimes there is no 'right answer' to be coded for. For example, it might seem straightforward to code a compliment as something that is nice to say. However, individuals with low self-esteem tend to interpret compliments as an inauthentic gesture from people who feel obligated to be nice, and they often tend to feel worse afterwards[45]. Because the same person might respond differently to the same text in different contexts, expert evaluations or annotations for subtle psychological constructs are not always robust. Additional sources of evaluation might be desirable.

Impact evaluation. Impact evaluations assess the LLM's effect on or utility for the intended user or consumer of the text (Box 3). For example, an LLM could be asked to generate a persuasive argument about the death penalty. Expert evaluation would entail social psychologists coding the persuasive qualities of an LLM-generated argument; impact evaluation would ask lay participants to report their attitudes before and after reading the argument. The latter method – focusing on persuasive impact – is a better way to assess a construct such as persuasiveness, as it directly determines whether the outcome (persuasion) was achieved, without relying on subjective and contentious criteria for persuasiveness[46]. Expert evaluation is currently the most common approach to evaluating persuasiveness in LLMs[47], pointing to the need for more impact evaluations.

One drawback of impact evaluation is that it can be theoretically ambiguous when LLM-generated text fails to change downstream outcomes: did it fail because the LLM output failed to communicate the psychological construct of interest effectively, or because the construct does not affect outcomes? Thus, although impact evaluation is more likely to serve as 'ground truth' for the effectiveness of language intended to reflect a given psychological construct, it is not informative about whether or to what extent the construct was embodied in that language.

Recommendations for evaluating LLMs. Because expert evaluation and impact evaluation offer complementary advantages and disadvantages, we recommend using both when possible. Expert evaluation methods can be used first, akin to a manipulation check or a measure of construct validity. Next, the scored text that matches expert judgements can be used in an impact evaluation study that attempts to measure the intended effects on third-party participants, akin to a measure of predictive or external validity.

This process can be iterative, in that impact evaluations can be used to inform the next round of expert evaluation and vice versa. For example, suppose a panel of experts identified principles for the best reappraisal techniques (techniques for changing one's thinking to change one's emotional response[48]). They could score a corpus of text along theoretically defined key dimensions, such as reframing all-or-nothing thinking about stressors in terms of temporary, controllable factors. Then they could fine-tune or prompt-tune an LLM to conform to these dimensions. An impact evaluation could assess whether LLM-generated text higher on these dimensions improved mental health or coping for real human participants. With the results of impact evaluations in hand, the panel could re-convene and debate new hypotheses about the validity of their theories and scoring guidelines, and repeat the impact evaluation with text generated by the newly updated LLM. The result would be a corpus of human and AI-generated text with known theoretical and impact qualities, which could power future automated scoring algorithms. This process could be followed for major constructs

Glossary

Backpropagation
An algorithmic technique that utilizes a reverse pass through the network to calculate the contribution of each parameter to the prediction error and adjust them accordingly to improve performance.

Bag-of-words methods
A text representation technique that counts the frequency of words in a document, disregarding grammar and word order, such as the Linguistic Inquiry and Word Count algorithm.

Generative pre-trained transformer
A family of large language models developed by OpenAI and usually trained on massive datasets to generate contextually coherent text.

Machine learning
A subset of artificial intelligence that involves teaching computers to learn patterns and make decisions from data without explicit programming.

Neural network
A computational model inspired by the structure and function of biological neural networks used for tasks such as pattern recognition, classification and prediction.

Training data
The dataset used to train a machine learning model, consisting of input–output pairs that help the model to learn the underlying patterns and relationships.

in each of psychology's subfields (presumably after a period of building consensus around which constructs are the highest priorities).

Bias in LLMs

LLM-generated text could perpetuate stereotypes and cause harm to vulnerable populations. Because LLMs learn from patterns in natural language training data, they reproduce biases in those data[49], possibly leading to representational and allocational harms[50,51]. Representational harms arise when the LLM represents some social groups in a less favourable light than others, demeans them, or fails to recognize their existence altogether. Such harms can be caused by sampling biases (for example, when privileged groups' data are over-represented) and data-processing biases (for example, when annotators from majority groups misunderstand minority groups). Allocational harms arise when AI algorithms differentially allocate resources (such as loans) or opportunities (such as therapy) to different social groups owing to historically biased decision patterns represented in the data, such as biased diagnoses or biased assignment to therapy treatment.

One pernicious bias in LLMs stems from the use of online text data to train them. Algorithms for generating and presenting text to people online through social media and news platforms are subject to both representational and allocational harms because they are driven by predictions about which text will engage targeted consumers and generate advertising or sales revenue[52]. Consequently, the 'average' sentence or utterance in the web data used to train LLMs is not a random sample of all sentences or utterances, but instead often reflects a previous algorithm's output. LLMs are already being used for copywriting online without transparency about whether it was written by a human or a machine. That text will probably be used to train subsequent generations of LLMs. Thus, without corrective action many biases might become more entrenched and more difficult to detect over time.

Using LLMs to study bias. Psychologists might be interested in studying LLMs' biases to gain insight into human bias. However, bias-probing is made difficult by censorship 'guardrails' that have been placed on most LLMs by their developers. For example, a typical LLM will not directly answer a question such as "are white people smart?". Thus, indirect methods to check for bias are needed instead. One method for interrogating bias in LLMs (often used in audit studies of real-world bias[53]) is manipulating pronouns and names associated with different genders, races and ethnicities in the input and assessing whether there are systematic differences in the output[29,54].

One limitation of this audit approach is that it requires explicit specification of demographic attributes, which makes it difficult to observe unknown biases or biases related to intersectionality. In addition, as such tests become increasingly common, LLMs developers might implement further censorship guardrails (via post-processing the output or tuning the model) to prevent specific problematic outputs in a way that addresses the symptoms, rather than the underlying bias in the data. These guardrails could falsely lead researchers to conclude that the LLM is not biased[55]. It is therefore a high priority to make censorship algorithms transparent and to develop bias-testing protocols that go beyond testing obvious stereotypes to also examine the pernicious, pervasive and subtle ways that bias manifests in culture[56,57]. Any such bias-probing will be further complicated by a lack of transparency over whether the language used to fit the LLM was itself generated by an LLM or by a human. Thus, even if better protocols are developed, it would not be obvious in many cases whether the researcher was studying human bias or algorithm-mediated bias[52].

Recommendations for mitigating bias in LLMs. One key approach to using LLMs ethically in psychological research despite their potential for bias is to involve people who are representative of the population that the LLM is meant to serve (for example, teachers and underserved students for a college advice tool) at every stage of the research process (research design, data annotation and model evaluation). Another way to mitigate bias is to take an 'AI-in-the-loop, humans in charge' approach[58]: instead of allowing the LLM to interact directly with vulnerable populations (for example, struggling students), human experts (for example, teachers) are in control of the interaction and can accept or reject suggestions from the LLM.

As LLMs become better at simulating different types of behaviour, they could be used to recognize and diagnose undesirable biases in the content they produce as well as text written by humans. For example, researchers are developing approaches for LLMs to self-diagnose[59] and self-criticize based on principles written by people[60] to de-bias their output. These algorithms could also be applied to mitigate human bias as a form of bias training. For example, participants could receive constructive feedback on their responses to scenarios that have the potential for bias. Given that many training programmes to combat implicit bias do not have a measurable benefit[61,62], LLM-guided explicit-bias training might be a promising avenue for research. In general, it will be important to develop clear processes for quantifying and reducing bias that are not solely driven by LLMs themselves.

The 'black box' of LLM output

LLM outputs are straightforward. As shown in Fig. 1, LLMs can be asked which emotions are present in a sentence, and they can provide an answer in plain language, just as a participant in a psychology study can. LLMs can also provide readily interpretable outputs for context-dependent, high-inference constructs, such as belonging (Box 2) or suicide risk[63]. This is not possible with outputs from conventional text-analysis methods, such as word-counting and topic modelling algorithms, which must be translated into a psychological construct by a researcher. For example, if a Linguistic Inquiry and Word Count model scores a text high on 'cognitive processing words', the researcher still needs to map 'cognitive processing' onto constructs such as anxiety or depression. Similarly, interpreting topics from a topic model is subjective[64], and the topics might not map onto psychological constructs of interest.

The interpretability of LLM output increases access to these tools for non-technical users and enables powerful applications, such as coaching tools for teachers, therapists and peers[65–67] or for students who are writing self-persuasion essays as part of 'wise' psychological interventions[68–71]. At the same time, it is difficult to interpret the processes by which LLMs reach their responses[45,72,73]. This is a problem because psychologists typically care about mechanisms. It is usually not enough to know that a sentence expresses anxiety; psychologists want to know why and how the language implies anxiety. Interrogating LLMs to identify the linguistic predictors or rationales of their predictions would inform theories of psychological constructs. However, it is not clear how to assign meaning to the billions of parameters in LLMs.

One widely used practice for interpreting LLMs is called probing[74]. In probing, researchers seek evidence on how the matrices underlying a model's inferences align with linguistic features that influence the model's outputs. To do this, researchers first need to identify linguistic features relevant to their psychological construct (for example, negation, use of first person, or use of cognitive processing words from the Linguistic Inquiry and Word Count lexicon) and manipulate the input along these features to test their effect on the output. For example, researchers can create a large number of input sentence pairs that are only distinguished by the presence of negation on verbs (for example, 'I care' versus 'I don't care'). They can then compare the models' outputs to the negated and non-negated sentences to determine whether the LLM takes negation into consideration when making its predictions, and if so, which values in the vector are most correlated with negation.

Although researchers have tried directly asking LLMs to explain their predictions[48,75,76] doing so creates a circular problem[45]. LLMs are not able to reason about their outputs, so examining their black box with yet another black box explanation is hardly a reliable approach.

With the uptake of LLMs in social science research, we anticipate that more tools to make LLMs' output interpretable to psychologists will be developed. We might also see newer models that more readily lend themselves to interpretation than existing LLMs. At this stage, we recommend complementing LLMs with traditional lexical analysis to connect interpretable outputs to interpretable processes.

Using LLMs as a reproducible tool

LLMs are stochastic (non-deterministic) models, meaning that they can generate different responses to the same prompt. LLMs are trained over natural language text and learn the distributions of possible words. There is therefore no single correct answer to how a given sentence should continue and end. The variability in LLM output is controlled by a temperature parameter that determines how it samples from the distribution of possible words. A temperature of 0 would have the model always pick the most likely words (the maximum a posteriori estimate), whereas positive temperatures allow the model to sample more widely from the distribution, which results in more linguistically (and possibly psychologically) diverse outputs.

This stochasticity is similar to how the same human research participant might respond differently when presented with the same prompt – there is a true 'signal', but there is also variability and noise. To address this variability, psychologists do not typically survey one participant; they survey many and then compute statistics on the aggregate. Increasingly, psychologists also randomly vary the stimuli in their studies[77]. In both cases, psychologists incorporate uncertainty intervals into their conclusions. Likewise, statisticians conducting simulation studies re-sample from their data generating process and incorporate that uncertainty into their analyses.

Our recommendation is therefore to acknowledge and incorporate the stochasticity of LLM outputs into analyses, similarly to how researchers account for other forms of stochasticity in psychological research. For example, researchers can use LLMs to generate multiple responses for any prompt and then save and publish (in supplementary material) both the parameters and outputs. This approach will aid reproducibility and also lead to more generalizable conclusions. We note that although one could in principle eliminate stochasticity (for example by setting temperature to 0), this might lead to missing insights that come from the diversity of the language in the model, and so we do not recommend this approach.

Outlook

The introduction of GPT (and related LLMs) is a revolutionary advancement in language-related technology that has profound implications for psychology. Massive amounts of language data can now be harnessed to help psychologists to develop better measures, to perform larger and perhaps more definitive experiments, and to develop and test ways of enhancing practice in areas that are critical to society, such as education and mental health.

Armed with an understanding of the conceptual foundations and concerns related to the use of LLMs in psychology, we turn to three important investments the field can make to realize the full potential of LLMs. Of course, it is unlikely that any single researcher could make these investments on their own. Thus, we foresee a need for collaborations across individual research groups, professional organizations (such as the Association for Psychological Science or the American Psychological Association), private industry and the funding community.

First, a priority is to invest in keystone datasets. These are large natural language databases that meet the highest standards for both expert and impact evaluation. Keystone datasets must represent populations and psychological constructs of interest; must rely on precise definitions of these psychological constructs (possibly including expert annotations); and must be linked to psychologically important outcomes, such as real-world behaviour, mindset, health and mental health, or academic and career trajectory.

For example, the Texas Mindset Initiative is creating a large dataset of 1,000 Texas secondary school teachers' natural language when talking with students. Teachers' transcripts can be annotated with expertise in different constructs, such as growth mindset (the belief that intellectual abilities can be developed), emotional reappraisal (novel ways of thinking that lead to decreased negative emotions) or autonomy support (whether teachers refrain from controlling language). These data could then be empirically linked with teachers' reports of their intentions to use these concepts, with students' ratings of classroom culture, and with students' academic performance over time.

Triangulating across these sources of data will enable discoveries. For example, researchers might learn what teachers can say to a struggling student to support their mindset, to help them to cope with stress and to help them to learn. These findings could then yield opportunities for fine-tuning LLMs to measure or recommend language that matches the constructs of interest, which could in turn power educational technologies such as virtual tutors or coaching tools for teachers. If similar datasets are collected in other educational contexts, for example in a different region, country or for a different developmental age group, these datasets could collectively shed light on the culture-dependent and age-dependent context-specificity of LLM-generated psychological language in teaching contexts.

Similar efforts could be undertaken by scholars interested in other constructs or policy spaces. The availability of online therapy platforms[66,78,79], body cameras on police officers[80] and recorded Zoom meetings offers much new material for keystone datasets in clinical psychology, policing and management. We envision a field in which experts have reached a consensus on a corpus of text that embodies major constructs in the field, which can be used to fine-tune LLMs and unlock opportunities for at-scale natural language analyses.

Second, the success of LLMs so far comes from the benchmarking culture in computer science[81]. Benchmarking occurs when model developers refine their software to optimize performance on a set of defined tasks and datasets[82,83]. However, existing forms of benchmarking might not be appropriate for many psychological use cases owing to the highly context-dependent nature of psychological constructs and intersubjectivity issues. Thus, a new, psychologically defined way of benchmarking LLMs could help to facilitate the development of safe and transparent algorithms that advance the field. Creating such benchmarks depends on the availability of keystone datasets and requires a consensus on how to define and operationalize psychological constructs. Finally, the field needs to create standards and procedures for safety and transparency, which might involve sharing models with other scientists for further evaluation. Organizing bodies (such as professional societies and funding agencies) could drive this consensus process.

Importantly, the privacy and confidentiality of language data will need to be accounted for in benchmarking practices. This is especially important when working with data from vulnerable populations. Instead of sharing their original data publicly, psychologists might share anonymized or modified versions of their data where randomized simulations are used to prevent reverse identification. There is a large and growing body of research and policy concerning AI and the use of private data, and the role of AI in perpetuating biases, which will be critical for the future of LLM-powered research[84–88].

Third, as psychologists increasingly see the potential for LLMs to influence their work, it will be important to ensure that access to these tools remains equitable. There are still only a handful of LLMs, because fitting each one takes an immense amount of resources. For example, the basic research and model fitting required to create ChatGPT cost US $3 billion[89]. Even the smaller, less costly LLMs can take months to train with the most advanced cloud computing resources. Because academia is unlikely to compete with the for-profit sector for the development of multi-billion-dollar models, researchers should invest in developing cost-effective LLMs whose performance is as close as possible to expensive, paywall-protected and non-downloadable models. It will be important to ensure that academics can access discounted or free research licences for using LLMs. To further lower barriers to entry, tutorials, optimal defaults and vetted data-processing pipelines should be created. Investments in time-shared research infrastructure, including access to servers or cloud computing, can ensure that a researcher's funding does not prohibit them from conducting AI-informed science. Moreover, working with LLMs responsibly might require a level of technical expertise that is not typically taught during psychology methods training. Thus, a new generation of text-analysis psychology methods courses and more training in programming languages such as Python might be needed.

Finally, we see many opportunities for the next generation of open science collaboration and team science to contribute to the future of LLMs in psychology. Psychological scientists and computational linguists can learn from one another and could contribute to aspects of keystone data collection and model fine-tuning. Interdisciplinary team science has the potential for contributions to both fields: particularly interesting use cases might be introduced to the computational linguistics literature from psychology, whereas the technical advances developed by computational linguists could empower new study designs and methods for psychology.

Published online: 13 October 2023

References

1. Pennebaker, J. W., Mehl, M. R. & Niederhoffer, K. G. Psychological aspects of natural language use: our words, our selves. *Annu. Rev. Psychol.* **54**, 547–577 (2003).
2. Stone, P. J., Dunphy, D. C. & Smith, M. S. *The General Inquirer: A Computer Approach to Content Analysis* Vol. 651 (MIT Press, 1966).
3. Landauer, T. K., Foltz, P. W. & Laham, D. An introduction to latent semantic analysis. *Discourse Process.* **25**, 259–284 (1998).
4. Landauer, T. K. & Dumais, S. T. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **104**, 211–240 (1997).
5. Choudhury, M. D., Gamon, M., Counts, S. & Horvitz, E. Predicting depression via social media. In *Proc. Int. AAAI Conf. Web Social Media* **7**, 128–137 (2013).
6. Eichstaedt, J. C. et al. Psychological language on Twitter predicts county-level heart disease mortality. *Psychol. Sci.* **26**, 159–169 (2015).
7. Boyd, R. L. & Schwartz, H. A. Natural language analysis and the psychology of verbal behavior: the past, present, and future states of the field. *J. Lang. Soc. Psychol.* **40**, 21–41 (2021).
8. Bhatia, S. & Aka, A. Cognitive modeling with representations from large-scale digital data. *Curr. Dir. Psychol. Sci.* **31**, 207–214 (2022).
9. Boyd, R., Ashokkumar, A., Seraj, S. & Pennebaker, J. *The Development and Psychometric Properties of LIWC-22* (Univ. Texas at Austin, 2022).
10. Blei, D., Ng, A. & Jordan, M. Latent Dirichlet allocation. In *Advances in Neural Information Processing Systems* (eds Dietterich, T., Becker, S. & Ghahramani, Z.) Vol. 14 (MIT Press, 2001).
11. Brown, T. et al. in *Advances in Neural Information Processing Systems* (eds Larochelle, H. et al.) Vol. 331, 877–1901 (Curran Associates, Inc., 2020).
12. OpenAI. GPT-4 technical report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.08774 (2023).
13. Collins, E. & Ghahramani, Z. LaMDA: our breakthrough conversation technology. *Google* https://blog.google/technology/ai/lamda/ (2021).
14. Wittgenstein, L. *Tractatus Logico-Philosophicus* (Edusp, 1994).
15. Wallace, J. Only in the context of a sentence do words have any meaning. *Midw. Stud. Phil.* **2**, 144–164 (1977).
16. Eliot, L. People are eagerly consulting generative AI ChatGPT for mental health advice, stressing out AI ethics and AI law. *Forbes* https://www.forbes.com/sites/lanceeliot/2023/01/01/people-are-eagerly-consulting-generative-ai-chatgpt-for-mental-health-advice-stressing-out-ai-ethics-and-ai-law/ (2023).
17. ChatGPT used by teachers more than students, new survey from Walton Family Foundation Finds. *Walton Family Foundation* https://www.waltonfamilyfoundation.org/chatgpt-used-by-teachers-more-than-students-new-survey-from-walton-family-foundation-finds (2023).
18. Ziegler, D. M. et al. Fine-tuning language models from human preferences. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1909.08593 (2020).
19. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
20. Weiss, M., Brock, T., Sommo, C., Rudd, T. & Turner, M. C. Serving community college students on probation: four-year findings from Chaffey College's Opening Doors Program. *MDRC* https://eric.ed.gov/?id=ED526395 (2011).
21. Crum, A. J., Akinola, M., Martin, A. & Fath, S. The role of stress mindset in shaping cognitive, emotional, and physiological responses to challenging and threatening stress. *Anxiety Stress Coping* **30**, 379–395 (2017).
22. Yeager, D. S. et al. A synergistic mindsets intervention protects adolescents from stress. *Nature* **607**, 512–520 (2022).
23. Crum, A. J., Salovey, P. & Achor, S. Rethinking stress: the role of mindsets in determining the stress response. *J. Pers. Soc. Psychol.* **104**, 716–733 (2013).
24. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? In *Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency* 610–623 (ACM, 2021).
25. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. *Proc. Natl Acad. Sci.* **120**, e2218523120 (2023).
26. Landi, H. Doximity rolls out beta version of ChatGPT tool for docs aiming to streamline administrative paperwork. *Fierce Healthcare* https://www.fiercehealthcare.com/health-tech/doximity-rolls-out-beta-version-chatgpt-tool-docs-aiming-streamline-administrative (2023).
27. Liu, X. et al. P-Tuning: prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proc. 60th Ann. Meet. Assoc. Computational Linguistics* Vol. 2, 61–68 (Association for Computational Linguistics, 2022).
28. Liu, P. et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **55**, 1–35 (2023).
29. Argyle, L. et al. Out of one, many: using language models to simulate human samples. *Polit. Anal.* **31**, 337–351 (2023).
30. Plaza-del-Arco, F. M., Martín-Valdivia, M.-T. & Klinger, R. Natural language inference prompts for zero-shot emotion classification in text across corpora. In *Proc. 29th Int. Conf. on Computational Linguistics* 6805–6817 (International Committee on Computational Linguistics, 2022).
31. Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B. & Yang, Q. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proc. 2023 CHI Conf. on Human Factors in Computing Systems* 1–21 (Association for Computing Machinery, 2023).
32. Park, J. S. et al. Social simulacra: creating populated prototypes for social computing systems. In *35th Ann. ACM Symp. on User Interface Software and Technology* 1–18 (Association for Computing Machinery, 2022).
33. Aher, G. V., Arriaga, R. I. & Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *Proc. 40th Int. Conf. on Machine Learning* 337–371 (PMLR, 2023).
34. Mahowald, K. et al. Dissociating language and thought in large language models: a cognitive perspective. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2301.06627 (2023).
35. Trott, S., Jones, C., Chang, T., Michaelov, J. & Bergen, B. Do large language models know what humans know? *Cogn. Sci.* **47**, e13309 (2023).
36. Sap, M., Le Bras, R., Fried, D. & Choi, Y. Neural theory-of-mind? On the limits of social intelligence in large LMs. In *Proc. 2022 Conf. on Empirical Methods in Natural Language Processing* 3762–3780 (Association for Computational Linguistics, 2022).
37. Marcus, G. & Davis, E. GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review* https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/ (2020).

38. Marcus, G. & Davis, E. Large language models like ChatGPT say the darnedest things. *The Road to AI We Can Trust* https://garymarcus.substack.com/p/large-language-models-like-chatgpt (2023).
39. OpenAI. *GPT-4 Technical Report* (2023).
40. Novikova, J., Dušek, O., Curry, A. C. & Rieser, V. Why we need new evaluation metrics for NLG. In *Proc. 2017 Conf. on Empirical Methods in Natural Language Processing* 2231–2240 (2017).
41. Luo, F. et al. Towards fine-grained text sentiment transfer. In *Proc. 57th Ann. Meet. Assoc. Computational Linguistics* 2013–2022 (Association for Computational Linguistics, 2019).
42. Lord, S. P., Sheng, E., Imel, Z. E., Baer, J. & Atkins, D. C. More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client. *Behav. Ther.* **46**, 296–303 (2015).
43. Schuetz, A. Scheler's theory of intersubjectivity and the general thesis of the alter ego. *Phil. Phenomenol. Res.* **2**, 323–347 (1942).
44. Fiske, S. T. Interpersonal stratification: status, power, and subordination. In *Handbook of Social Psychology* 5th edn, Vol. 2 (eds Fiske, S. T., Gilbert, D. T. & Lindzey, G.) 941–982 (John Wiley & Sons, 2010).
45. Lai, V., Zhang, Y., Chen, C., Liao, Q. V. & Tan, C. Selective explanations: leveraging human input to align explainable AI. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2301.09656 (2023).
46. Petty, R. E. & Cacioppo, J. T. The elaboration likelihood model of persuasion. In *Advances in Experimental Social Psychology* Vol. 19 (ed. Berkowitz, L.) 123–205 (Academic Press, 1986).
47. Karinshak, E., Liu, S. X., Park, J. S. & Hancock, J. T. Working with AI to persuade: examining a large language model's ability to generate pro-vaccination messages. *Proc. ACM Hum. Comput. Interact.* **7**, 116 (2023).
48. Gross, J. J. Antecedent- and response-focused emotion regulation: divergent consequences for experience, expression, and physiology. *J. Pers. Soc. Psychol.* **74**, 224–237 (1998).
49. Harris, C., Halevy, M., Howard, A., Bruckman, A. & Yang, D. Exploring the role of grammar and word choice in bias toward African American English (AAE) in hate speech classification. In *2022 ACM Conf. on Fairness, Accountability, and Transparency* 789–798 (Association for Computing Machinery, 2022).
50. Barocas, S., Hardt, M. & Narayanan, A. *Fairness and Machine Learning: Limitations and Opportunities* https://fairmlbook.org/ (MIT Press, in the press).
51. Blodgett, S. L., Barocas, S., Daumé III, H. & Wallach, H. Language (technology) is power: a critical survey of 'bias' in NLP. Preprint at *arXiv* http://arxiv.org/abs/2005.14050 (2020).
52. Brady, W. J., Jackson, J. C., Lindström, B. & Crockett, M. Algorithm-mediated social learning in online social networks. Preprint at *OSFPreprints* https://doi.org/10.31219/osf.io/yw5ah (2023).
53. Gaddis, S. M. An introduction to audit studies in the social sciences. In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance* (ed. Gaddis, S. M.) 3–44 (Springer International Publishing, 2018).
54. Lucy, L. & Bamman, D. Gender and representation bias in GPT-3 generated stories. In *Proc. Third Worksh. on Narrative Understanding* 48–55 (Association for Computational Linguistics, 2021).
55. Gonen, H. & Goldberg, Y. Lipstick on a pig: debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1, 609–614 (Association for Computational Linguistics, 2019).
56. Cheryan, S. & Markus, H. R. Masculine defaults: identifying and mitigating hidden cultural biases. *Psychol. Rev*. **127**, 1022–1052 (2020).
57. Walton, G. M., Murphy, M. C. & Ryan, A. M. Stereotype threat in organizations: implications for equity and performance. *Annu. Rev. Organ. Psychol. Organ. Behav.* **2**, 523–550 (2015).
58. Monarch, R. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI* (Simon and Schuster, 2021).
59. Schick, T., Udupa, S. & Schütze, H. Self-diagnosis and self-debiasing: a proposal for reducing corpus-based bias in NLP. *Trans. Assoc. Comput. Linguist.* **9**, 1408–1424 (2021).
60. Bai, Y. et al. Constitutional AI: harmlessness from AI feedback. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2212.08073 (2022).
61. Chang, E. H. et al. The mixed effects of online diversity training. *Proc. Natl Acad. Sci.* **116**, 7778–7783 (2019).
62. Lai, C. K. et al. Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *J. Exp. Psychol. Gen.* **143**, 1765–1785 (2014).
63. Allen, N. B., Nelson, B. W., Brent, D. & Auerbach, R. P. Short-term prediction of suicidal thoughts and behaviors in adolescents: can recent developments in technology and computational science provide a breakthrough? *J. Affect. Disord.* **250**, 163–169 (2019).
64. Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L. & Blei, D. M. Reading tea leaves: how humans interpret topic models. *Adv. Neural Inf. Process. Syst.* **22**, 288–296 (2009).
65. Demszky, D., Liu, J., Hill, H. C., Jurafsky, D. & Piech, C. Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educ. Eval. Policy Anal.* https://doi.org/10.3102/01623737231169270 (2023).
66. Shah, R. S. et al. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *Proc. ACM Hum. Comput. Interact.* **6**, 1–24 (2022).
67. Demszky, D. & Liu, J. M-Powering teachers: natural language processing powered feedback improves 1:1 instruction and student outcomes. In *Proc. Tenth ACM Conf. on Learning @ Scale* 59–69 (Association for Computing Machinery, 2023).
68. Aronson, E. The power of self-persuasion. *Am. Psychol.* **54**, 875–884 (1999).
69. Walton, G. M. & Wilson, T. D. Wise interventions: psychological remedies for social and personal problems. *Psychol. Rev.* **125**, 617–655 (2018).
70. Walton, G. M. & Cohen, G. L. A brief social-belonging intervention improves academic and health outcomes of minority students. *Science* **331**, 1447–1451 (2011).
71. Yeager, D. S. et al. A national experiment reveals where a growth mindset improves achievement. *Nature* **573**, 364–369 (2019).
72. Wang, P., Chan, A., Ilievski, F., Chen, M. & Ren, X. PINTO: faithful language reasoning using prompt-generated rationales. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2211.01562 (2022).
73. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should I trust you?': explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* 1135–1144 (ACM, 2016).
74. Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U. & Levy, O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl Acad. Sci.* **117**, 30046–30054 (2020).
75. Simonsohn, U., Nelson, L. D. & Simmons, J. P. P-curve: a key to the file-drawer. *J. Exp. Psychol. Gen.* **143**, 534–547 (2013).
76. Messick, S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* **50**, 741–749 (1995).
77. Judd, C. M., Westfall, J. & Kenny, D. A. Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *J. Pers. Soc. Psychol.* **103**, 54–69 (2012).
78. Wang, T. et al. Metrics for peer counseling: triangulating success outcomes for online therapy platforms. In *Proc. 2023 CHI Conf. on Human Factors in Computing Systems* 1–17 (ACM, 2023).
79. Nook, E. C., Hull, T. D., Nock, M. K. & Somerville, L. H. Linguistic measures of psychological distance track symptom levels and treatment outcomes in a large set of psychotherapy transcripts. *Proc. Natl Acad. Sci.* **119**, e2114737119 (2022).
80. Voigt, R. et al. Language from police body camera footage shows racial disparities in officer respect. *Proc. Natl Acad. Sci.* **114**, 6521–6526 (2017).
81. Paullada, A., Raji, I. D., Bender, E. M., Denton, E. & Hanna, A. Data and its (dis)contents: a survey of dataset development and use in machine learning research. *Patterns* **2**, 100336 (2021).
82. Wang, A. et al. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In *Proc. 2018 EMNLP Worksh. BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* 353–355 (Association for Computational Linguistics, 2018).
83. Gehrmann, S. et al. The GEM benchmark: natural language generation, its evaluation and metrics. In *Proc. 1st Worksh. on Natural Language Generation, Evaluation and Metrics (GEM 2021)* 96–120 (Association for Computational Linguistics, 2021).
84. Birhane, A. & Guest, O. Towards decolonising computational sciences. Preprint at *arXiv* http://arxiv.org/abs/2009.14258 (2020).
85. Birhane, A. Algorithmic injustice: a relational ethics approach. *Patterns* **2**, 100205 (2021).
86. Erscoi, L., Kleinherenbrink, A. & Guest, O. Pygmalion displacement: when humanising AI dehumanises women. Preprint at *SocArXiv* https://doi.org/10.31235/osf.io/jqxb6 (2023).
87. Guest, O. & Martin, A. E. On logical inference over brains, behaviour, and artificial neural networks. *Comput. Brain Behav.* **6**, 213–227 (2023).
88. Llorens, A. et al. Gender bias in academia: a lifetime problem that needs solutions. *Neuron* **109**, 2047–2074 (2021).
89. Metz, C. & Weise, K. Microsoft bets big on the creator of ChatGPT in race to dominate AI. *The New York Times* (12 January 2023).
90. Tesfagergish, S. G., Kapočiūtė-Dzikienė, J. & Damaševičius, R. Zero-shot emotion detection for semi-supervised sentiment analysis using sentence transformers and ensemble learning. *Appl. Sci.* **12**, 8662 (2022).
91. ElSherief, M. et al. Latent hatred: a benchmark for understanding implicit hate speech. In *Proc. 2021 Conf. on Empirical Methods in Natural Language Processing* 345–363 (Association for Computational Linguistics, 2021).
92. Pryzant, R. et al. Automatically neutralizing subjective bias in text. *Proc. AAAI Conf. Artif. Intell.* **34**, 480–489 (2020).
93. Ophir, Y., Tikochinski, R., Asterhan, C. S. C., Sisso, I. & Reichart, R. Deep neural networks detect suicide risk from textual Facebook posts. *Sci. Rep.* **10**, 16685 (2020).
94. Basta, C., Costa-jussà, M. R. & Casas, N. Evaluating the underlying gender bias in contextualized word embeddings. In *Proc. First Worksh. on Gender Bias in Natural Language Processing* 33–39 (Association for Computational Linguistics, 2019).
95. Ashokkumar, A. & Pennebaker, J. W. Social media conversations reveal large psychological shifts caused by COVID-19's onset across US cities. *Sci. Adv.* **7**, eabg7843 (2021).
96. Rathje, S. et al. GPT is an effective tool for multilingual psychological text analysis. Preprint at *PsyArXiv* https://psyarxiv.com/sekf5/ (2023).
97. Seraj, S., Blackburn, K. G. & Pennebaker, J. W. Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup. *Proc. Natl Acad. Sci.* **118**, e2017154118 (2021).
98. Sap, M. et al. Quantifying the narrative flow of imagined versus autobiographical stories. *Proc. Natl Acad. Sci.* **119**, e2211715119 (2022).

99. Michelmann, S., Kumar, M., Norman, K. A. & Toneva, M. Large language models can segment narrative events similarly to humans. Preprint at *arXiv* http://arxiv.org/abs/2301.10297 (2023).
100. Zhang, S., She, J. S., Gerstenberg, T. & Rose, D. You are what you're for: essentialist categorization in large language models. In *Proc. Ann. Meet. Cognitive Science Society* Vol. 45 (2023).
101. Cimpian, A. & Salomon, E. The inherence heuristic: an intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behav. Brain Sci.* **37**, 461–480 (2014).
102. Portelance, E., Degen, J. & Frank, M. C. Predicting age of acquisition in early word learning using recurrent neural networks. In *Proc. Ann. Meet. Cognitive Science Society* (2020).
103. Westerveld, M. F. & Roberts, J. M. A. The oral narrative comprehension and production abilities of verbal preschoolers on the autism spectrum. *Lang. Speech Hear. Serv. Sch.* **48**, 260–272 (2017).
104. Siddaway, A. P., Wood, A. M. & Hedges, L. V. How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annu. Rev. Psychol.* **70**, 747–770 (2019).
105. Tipton, E., Pustejovsky, J. E. & Ahmadi, H. Current practices in meta-regression in psychology, education, and medicine. *Res. Synth. Meth.* **10**, 180–194 (2019).
106. Aher, G., Arriaga, R. I. & Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. Preprint at *arXiv* http://arxiv.org/abs/2208.10264 (2023).
107. Pennycook, G. et al. Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590–595 (2021).
108. Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T. & Van Bavel, J. J. An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *J. Exp. Psychol. Gen.* **148**, 1802–1813 (2019).
109. Milkman, K. L. et al. Megastudies improve the impact of applied behavioural science. *Nature* **600**, 478–483 (2021).
110. Stade, E. et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/cuzvr (2023).
111. Jacobs, J. et al. Promoting rich discussions in mathematics classrooms: using personalized, automated feedback to support reflection and instructional change. *Teach. Teach. Educ.* **112**, 103631 (2022).
112. Hunkins, N., Kelly, S. & D'Mello, S. "Beautiful work, you're rock stars!": teacher analytics to uncover discourse that supports or undermines student motivation, identity, and belonging in classrooms. In *LAK22: 12th Int. Learning Analytics and Knowledge Conf.* 230–238 (ACM, 2022).
113. Schwartz, H. A. et al. DLATK: Differential Language Analysis ToolKit. In *Proc. 2017 Conf. on Empirical Methods in Natural Language Processing: System Demonstrations* 55–60 (Association for Computational Linguistics, 2017).
114. Liu, Y. et al. RoBERTa: a robustly optimized BERT pretraining approach. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1907.11692 (2019).
115. Walton, G. M. et al. Where and with whom does a brief social-belonging intervention promote progress in college? *Science* **380**, 499–505 (2023).
116. Clapper, M. et al. Evaluating LLM's generation of growth-mindset supportive language in middle years math. Artificial intelligence in education. In *Proc. Worksh. on Equity, Diversity, and Inclusion in Educational Technology Research and Development 24th Int. Conf. on Artificial Intelligence in Education* (2023).
117. Hecht, C. A., Yeager, D. S., Dweck, C. S. & Murphy, M. C. Beliefs, affordances, and adolescent development: lessons from a decade of growth mindset interventions. *Adv. Child. Dev. Behav.* **61**, 169–197 (2021).
118. Hecht, C. A., Dweck, C. S., Murphy, M. C., Kroeper, K. M. & Yeager, D. S. Efficiently exploring the causal role of contextual moderators in behavioral science. *Proc. Natl Acad. Sci.* **120**, e2216315120 (2023).

Acknowledgements

This work was supported by the National Science Foundation under award numbers 1761179 and 2201928 (PI: D.S.Y.), by the National Institutes of Health under award numbers R01HD084772 (PI: D.S.Y.) and P2CHD042849 (Population Research Center), and by the William and Melinda Gates Foundation under awards INV-047751 and INV-004519 (PI: D.S.Y.). This work was also supported by an Advanced Research Fellowship from the Jacobs Foundation to D.S.Y., and the Institute for Human-Centered A.I. at Stanford University to J.C.E. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. The authors also thank C. Smith for creating the original version of the figures included with the original submission. The glossary definitions were generated by GPT-4 in May 2023 and edited by the authors.

Author contributions

Lead authors D.D., D.Y. and D.S.Y. (equal contributions, listed alphabetically) conceived the paper, outlined and wrote the first draft, guided the co-authoring process, provided critical edits, conceived and supervised the creation of the figures, boxes and tables, and finalized the submitted version of the manuscript. Senior authors C.S.D., J.J.G. and J.W.P. (listed alphabetically) assisted in the outlining, organization, and conceptualization of the manuscript, boxes and tables and provided multiple rounds of critical edits. J.C.E. assisted with outlining the paper, wrote first drafts of key sections and edited the draft. All other authors assisted with the empirical examples, the conceptualization of the key arguments and conclusions in the paper and provided critical edits.

Competing interests

J.W.P. is the CEO of Pennebaker Conglomerates, a company that sells natural language processing software and services. S.C. and L.L. are employees of Google LLC, which owns LLM technology. D.K.-C. was formerly an employee at Google LLC.

Additional information

Peer review information *Nature Reviews Psychology* thanks William Brady and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature America, Inc. 2023