Ethics and Information Technology (2024) 26:47
https://doi.org/10.1007/s10676-024-09778-2

ORIGINAL PAPER

Beyond transparency and explainability: on the need for adequate and contextualized user guidelines for LLM use

Kristian González Barman[1] · **Nathan Wood**[1] · **Pawel Pawlowski**[1]

Accepted: 16 May 2024 / Published online: 17 July 2024

Abstract

Large language models (LLMs) such as ChatGPT present immense opportunities, but without proper training for users (and potentially oversight), they carry risks of misuse as well. We argue that current approaches focusing predominantly on transparency and explainability fall short in addressing the diverse needs and concerns of various user groups. We highlight the limitations of existing methodologies and propose a framework anchored on user-centric guidelines. In particular, we argue that LLM users should be given guidelines on what tasks LLMs can do well and which they cannot, which tasks require further guidance or refinement by the user, and context-specific heuristics. We further argue that (some) users should be taught to refine and elaborate adequate prompts, be provided with good procedures for prompt iteration, and be taught efficient ways to verify outputs. We suggest that for users, shifting away from looking at the technology itself, but rather looking at the usage of it within contextualized sociotechnical systems, can help solve many issues related to LLMs. We further emphasize the role of real-world case studies in shaping these guidelines, ensuring they are grounded in practical, applicable strategies. Like any technology, risks of misuse can be managed through education, regulation, and responsible development.

Keywords Large language models (LLMs) · User guidelines · Explainable artificial intelligence (XAI) · AI ethics

Introduction

The rising impact of large language models (LLMs) like ChatGPT across various sectors has been nothing short of revolutionary, with these models being used more and more for an increasing variety of tasks, from streamlining mundane office tasks (Hadi et al., 2023a, 2023b) to pushing boundaries in scientific research (Boiko et al., 2023; Fan et al., 2023). The benefits are clear: LLMs allow humans to work more efficiently and with less time needed per task (Noy & Zhang, 2023), opening up new possibilities as a result. But, as with any technological leap, there are risks and challenges that must be addressed, especially relating to potential misuses and the need for better user understanding of these novel systems, their capabilities, and more importantly, their limitations.

Considering first their advantages, LLMs can be transformative to how we handle our daily workload, tackling everything from drafting emails to simplifying analyses, saving us time and effort. In science, LLMs are both speeding up the writing of research papers (Williams et al., 2023) and playing a role in conducting experiments and analyzing data (Inagaki et al., 2023; Jablonka et al., 2023; Xiao et al., 2023).

However, the more these models are used, the more we see the risks which attend that usage. There are widely recognized issues such as bias (Abid et al., 2021; Ferrara, 2023), inaccuracies (Bender et al., 2021), plagiarism (Lee et al., 2023), and the spreading of misinformation (Pan et al., 2023), but there are far more common and potentially impactful issues as well. In particular, everyday users of LLMs are apt to mistake what the systems are designed for, what they can do well, and what sorts of tasks are utterly unsuited to LLMs, leading to errors, misinformation, and misunderstanding, potentially across whole societies. Part of the problem is simply that most people do not understand

✉ Kristian González Barman
KristianCampbell.GonzalezBarman@ugent.be

Nathan Wood
nathan.wood@ugent.be

Pawel Pawlowski
pawel.pawlowski@ugent.be

[1] Centre for Logic and Philosophy of Science, Department of Philosophy and Moral Sciences, Ghent University, Blandijnberg 2, 9000 Ghent, Belgium

these tools, and also do not know what the tools can and can't do or how to use them properly. Underlying this former point is the fact that LLMs often come across as mysterious 'black boxes', where a prompt is introduced to get a response, without the user having any understanding of what's happening inside. And if the output looks good—e.g., reads like proper English and makes sense on the surface—people tend to use it even if it might not ultimately be correct.

One popular suggestion for addressing these issues is to increase the transparency and explainability of AI systems like large language models. In this article, we argue that this may not be the best or even the right sort of strategy regarding the use of LLMs. We argue that the focal point should be providing users with adequate guidelines such as specific statements regarding which (types of) tasks can be delegated to LLMs, which prompts are apt to lead to errors, misinformation, or unhelpful outputs, and what areas allow for collaborative use of LLMs in conjunction with human input and revision. The emphasis of such guidelines should not just (or even primarily) be about explaining LLMs or explaining how LLMs work, but rather about teaching people how to use them responsibly, ethically, and effectively, even if they do not (or cannot) understand the model or why it gave a certain output. More importantly, even when users are not properly 'taught' how to effectively or responsibly use LLMs, rudimentary guidelines can help everyday users to avoid common mistakes by simply providing them with a list of dos-and-don'ts for LLM use. As an example, one may think of using an LLM as analogous in some respects to driving a car; one does not need to be a mechanic who understands every detail about how the engine works in order to be a responsible driver. However, one should know the rules of the road and how to drive safely and effectively. The same is true for large language models (as well as many other opaque systems) (Wood, 2024). Understanding the theory behind them, or the reasons for why they fail is one thing, but knowing how to use them efficiently, effectively, or without causing harm is another.

While issues related to opacity, transparency, and explainability have been central points in the academic debates surrounding LLMs (and AI more generally), (Boge, 2022; Durán, 2021; Valentino & Freitas, 2022), it is not clear that explanations stemming from explainable AI (XAI) can improve the situation of the user in the case of LLMs. More specifically, we argue that XAI cannot properly address key challenges raised by LLMs and may even contribute to fostering neglect of an equally, if not more important topic when it comes to models used by the general public, namely good strategies and practices for everyday users. This includes teaching users the right way to use these tools, as well as informing them of not only the strengths, potential uses, and best practices for making use of them (e.g., prompt engineering strategies), but also the limitations and expected failings of these systems (Kasneci et al., 2023).

The *main aim* of this paper is to argue for the importance of LLM user guidelines to support reliable and responsible use of these technologies. In particular, to achieve reliable and responsible use of large language models, we argue that user guidelines are more effective than XAI, as users primarily seek practical guidance rather than explanations; users' concerns center around *how* to use LLMs effectively, rendering *why* explanations less useful. The paper further explores possible broad conceptions of what these guidelines should entail, and who should be administering them.

To be clear at the outset, we are not arguing that XAI is not useful (to some degree, it is likely that any sensible guideline's starting point should be XAI), but rather that focusing on explanations may be a suboptimal approach for users. This is because, even if good explanations (whether global or local) were available, providing them to users would not necessarily lead to reliable, efficient, and ethical use of LLMs. We therefore maintain that while XAI may help to elucidate the workings of LLMs, or might enable understanding the reasons for a certain output, it is not the most effective method for guiding users. Instead, practical guidelines should be emphasized, as they can fulfill many objectives similar to those aimed for by XAI without necessitating explanations. Thus, the reasons behind phenomena like LLM 'hallucinations' are less pertinent to users than simply knowing how to handle such occurrences effectively, especially as most users are arguably less concerned with understanding the underlying 'why' and are more interested in the practical how-to's and heuristics for responsible LLM use.

The key gain provided by this user-centric approach is the enhancement of user competence and confidence in utilizing LLMs responsibly and effectively. By focusing on practical guidelines, users are equipped with actionable knowledge, enabling them to better harness the potential of LLMs in various applications while reducing risks like biases and misinformation. This approach democratizes the use of advanced AI technologies, making them more accessible to a broader audience, irrespective of their technical background, thereby fostering a safer, more ethical, and more efficient use of LLMs.

The paper is structured as follows. Sect. "LLM use in education, the workplace, and expert advice" explores practical use cases of LLMs in contexts like education, the workplace, and expert seeking advice. This section highlights the need for policies, guidelines, and training tailored to different disciplines and domains where lack of reliability can have serious negative consequences. Sect. "Shortcomings and potential pitfalls of LLMs in practical applications" discusses the strengths and weaknesses of LLMs with an eye to potential pitfalls of using LLMs without proper guidelines.

Sect. "The challenges of XAI for LLMs" considers XAI as an approach to enable key desiderata in LLM use, arguing that it has several shortfalls and issues. Sect. "Charting a user-centric path: the case for tailored guidelines in a user-centric-approach (UCA)." advocates for a user-centric approach centered on the notion of guidelines and their strategic importance. Sect. "Meta criteria for guidelines" considers some meta-criteria for guidelines, considerations related to who should be administering said guidelines and how they should be implemented, and Sect. "Conclusion" concludes.

Preliminaries and clarifications

However, before moving onto the arguments, there are a few preliminary clarifications worth addressing. First, it is worth mentioning that to date there are a number of large language models available for use by everyday individuals, with LLMs such as OpenAI's GPT family, Gemini, Llama, and Claude becoming household names in the industry. While most current LLMs possess broad competence, some are better suited to particular tasks than others (Agarwal et al., 2024). In what follows, our focus is on broad aspects of LLMs at a general level, and as such, we will usually gloss over these particularities of model and version.

Second, LLMs are not just being utilized as standalone applications, but are also seeing potential incorporation into other tools such as search engines, document preparation systems, database manipulation software (e.g., tools such as Word, or Excel), or in custom programs that make use of an API. This potentially compounds issues of understandability for users, as it may become unclear what outputs are coming from an opaque LLM and which are built into more basic and interpretable programs. Moreover, embedding LLMs into other applications may complicate the presentation of guidelines for users, or create uncertainties concerning who is responsible for creating and implementing the guidelines in end-products. A full account of how best to tackle these issues is certainly necessary but is unfortunately beyond the scope of this article. However, the arguments developed here provide forceful grounds for exploring these issues more fully and addressing them in a user-centric manner as quickly as possible.

Third, as the arguments developed here focus on the limits of explanations in helping users to effectively and responsibly utilize opaque AI systems like LLMs, it will be useful to provide brief clarifications of what we mean by opacity and explainable AI (XAI). Roughly, the opacity or transparency of an AI system can be understood in terms of how well humans can perform a certain epistemic role (Humphreys, 2009). Opacity may arise due to a variety of factors, such industry secrecy, technical limitations of agents, or simple due to limitations in human cognition (Burrell, 2016). The purpose of explainable AI methods is to remedy some of the challenges posed by opacity. In particular, one of the main goals of XAI is to enable humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners (Arrieta et al., 2020). There are many approaches that go from trying to build interpretable models from the beginning (Rudin, 2019) to using statistical and computational methods to post hoc analyze a complex ML model or its outputs. This latter strategy might include figuring out input–output dependency relations, building surrogate models, or trying to understand smaller local mechanisms by which certain inputs turn into certain outputs (also known as mechanistic interpretability).

Finally, it is worth briefly specifying the nature of proposed guidelines we discuss in the final portions of this article. As mentioned above, our main aim is to argue that XAI is not entirely apt to solving some of the most fundamental issues relating to common uses of LLMs, and that instead we should be looking to institutional and human- or user-centric approaches. In particular, we argue that general user guidelines and best practices will provide a surer means for entrenching effective and responsible use of LLMs in a variety of situations. In exploring these points, we will propose certain desiderata for guidelines and also examine potential concrete guidelines as examples. However, the aim of this article is not to provide policy recommendations or design user interfaces. As such, all guidelines discussed here should be viewed as examples or first conceptions only.

LLM use in education, the workplace, and expert advice

The integration of large language models like GPT, Claude Llama, and Bard Bing Chat into educational settings presents a landscape of opportunities and challenges, as many universities are actively discovering. In response to these, some institutions are pioneering the establishment of guidelines for the use of LLMs, aiming to balance innovation with academic integrity.[1] They emphasize the importance of clear rules and proper citation of AI-generated content, highlighting a growing trend among universities to develop policies guiding AI tools' use in academic contexts.

In teaching, ChatGPT and similar LLMs offer significant potential benefits. They can enhance digital literacy and critical thinking, enabling students to analyze these tools'

[1] For example, Harvard, the University of California, Berkeley, and the University of Missouri have spearheaded efforts to codify guidelines on responsible and ethical use of LLMs within the university context. See https://provost.harvard.edu/guidelines-using-chatgpt-and-other-generative-ai-tools-harvard, https://ethics.berkeley.edu/privacy/appropriate-use-chatgpt-and-similar-ai-tools, https://oai.missouri.edu/chatgpt-artificial-intelligence-and-academic-integrity/.

strengths and limitations (Essel et al., 2024; Guo & Lee, 2023). Innovative pedagogical methods like "Think-Pair-ChatGPT-Share" utilize ChatGPT to foster deeper engagement with course materials (Yell, 2023). To prevent misuse, instructors can design assignments that demand specific references or personalization, underscoring the importance of original thought and the writing process. Importantly, instructors also can (and arguably should) design assignments which allow students to demonstrate for themselves the limits of these AI tools, showing the need for students to recognize these *as tools*, and ones which can be used well or poorly. For example, an instructor might assign students a text to analyze and then provide a short summary of the same text from ChatGPT (or similar) and, taking these together, require students to identify points or ideas which the LLM overlooked.

A key area where ChatGPT shines across domains is in automating tasks and providing feedback. Firstly, students have the opportunity to pose elementary inquiries about diverse technical concepts. LLMs can provide working definitions of such concepts as well as concrete examples. Thus, LLMs not only furnish responses but can also engage in subsequent questioning, offering a conversational learning dynamic.

From the educator's perspective, LLMs can be utilized to craft customized exercises augmented with automated feedback, and educators usually have the ability to amend insufficient or incorrect feedback the system might provide. Moreover, LLMs offer the potential to create novel teaching tools, such as quizzes, flashcards, and interactive games, thereby boosting student involvement and comprehension (Extance, 2023). ChatGPT can also act as a personal tutor, providing feedback on grammar, style, and argument consistency, supporting a tailored learning environment that adapts to each student's needs.

The flexibility of LLMs also allows for the tailoring of curricula to students with disabilities, creating an inclusive learning environment (Rakap, 2023; Choi, 2023). Teaching students to refine prompts and interact effectively with ChatGPT is crucial though, including instructing them on verifying information, acknowledging contributions from LLMs, and understanding how to use these tools optimally.

Scholarly research on LLMs in education provides a spectrum of views; researchers like Silva et al. (2023) and Lin (2023) recognize their potential in augmenting academic work, emphasizing ethical use and transparency. On the other hand, Dergaa et al. (2023) and Qadir (2023) raise concerns about authenticity and biases. Yan et al. (2023) and Yadav (2023) acknowledge the benefits of LLMs in automating educational tasks, but also highlight practical and ethical concerns. Fine Licht (2023), Kim et al. (2023), and Vidgof et al. (2023) further emphasize the need for transparent integration processes and guidelines in educational and business contexts.

However, the utilization of LLMs extends well beyond the educational realm, encompassing workspaces, elder care, mental health, and many other aspects of general well-being, highlighting the transformative potential of these emerging technologies. However, to fully capitalize on the capabilities of these advanced AI tools, it is crucial that users are instructed on how to properly engage with them, with strategies and comprehensive guidelines provided to ensure effective and responsible utilization.

In professional settings, LLMs are seeing increasing use, particularly in administrative tasks. Their proficiency in automating email responses and distilling key points from extensive meeting transcripts is revolutionizing time management and information dissemination. They also aid in the preliminary drafting of policies and assimilate vast amounts of information for informed decision-making. All in all, the ability of LLMs to streamline routine tasks is freeing employees to focus on more complex, creative work, enhancing brainstorming, spurring collaborative creativity, and fostering problem-solving. Maximizing these benefits, however, requires users to be skilled in crafting precise prompts and interpreting the responses appropriately. This underlines the necessity of basic training programs and elementary guidelines that not only enhance technical proficiency but also impart an understanding of the nuanced interactions users can and should have with AI.

LLMs also offer tailored solutions to specific organizational challenges. In customer service, ChatGPT can draft initial responses, later refined by humans, ensuring efficient yet personalized communication. In research and development, LLMs speed up project initiation and idea generation, thereby fostering innovation (Girotra et al., 2023). The effective deployment of LLMs in workplaces underscores their role as tools that complement human abilities, necessitating clear guidelines for their ethical and effective use.

A final area worth mention is elderly care and mental health, where LLMs can offer support by providing companionship and preliminary wellness advice (Fear & Gleber, 2023). However, it is imperative to educate users and provide basic guidelines and warnings regarding the limitations of LLMs in these sensitive areas, and to stress the importance of supplementing AI interactions with professional care.

Overall, the clear trend is for broader and more pointed use of LLMs across a host of domains. This can represent a positive development for individuals, organizations, and whole societies, with this new technology being leveraged for substantive gains of both an economic and human sort. However, as with any technology, LLMs can be used well or poorly, and in order for these tools to foster human development and flourishing, it is critical that we are attentive to not just their positive aspects, but also their shortcomings and

limitations. Most importantly, we must be vigilant in ensuring that users are using these technologies for their intended purposes, and that misuses of the system are minimized to the extent possible.[2]

Shortcomings and potential pitfalls of LLMs in practical applications

Though LLMs can significantly improve task automation and innovation across disparate fields, the various examples of LLM use in domains such as law, healthcare, and education present a mixed bag of successes and failures. These cases highlight the practical challenges in integrating LLMs into daily operations and decision-making processes. For instance, in the legal domain, while LLMs have been instrumental in processing large volumes of case files and legal documents, they have also faced criticism for oversimplifying complex legal reasoning or misinterpreting nuances in legal language (Sun, 2023).

The first major issue stems from an overestimation of LLMs' capabilities. Users, especially those not deeply familiar with AI models, may attribute too much intelligence, understanding, or capability to LLMs. Such overestimation can lead to unwarranted trust in the outputs of these models, without sufficient scrutiny or oversight. For instance, in sectors like healthcare or law, where the stakes are high, unquestioning reliance on LLMs for diagnosis or legal advice can lead to serious consequences, even when they are being used by experts in these fields. The critical fact for users to bear in mind is that these models, while powerful, should be used *as tools* for assistance rather than as ultimate decision-makers.

Overestimation of LLMs' capabilities can lead to a related problem when used by individuals who are not themselves experts in the domain within which they are prompting the system. For example, in legal or healthcare domains, use of LLMs by professionals entails a form of check on misuse of the system, as these experts are in a strong(er) position to recognize when the LLM is providing erroneous or potentially dangerous outputs. However, laypersons asking questions on these sorts of topics are particularly at risk of taking the LLMs' outputs at face value. Thus, the precise magnitude of risk involved in using an LLM for a given task may be both context- and user-dependent and show deep interrelations between these two factors.

Moreover, naïve use of LLMs work can lead to privacy and security issues, as users may unknowingly input sensitive or personal information into these models, not realizing that this data could be stored or used in ways that compromise privacy. Educating users and providing sufficient warnings and disclaimers about data security and privacy concerns associated with LLMs is crucial to prevent such breaches.

Similarly, the reliability of information provided by LLMs is a critical issue, as there have been instances of misinformation spreading due to erroneous outputs.[3] These erroneous outputs might simply be due to so-called 'hallucinations' (OpenAI, 2023), but they can also be due to a lack of contextualization; while LLMs are good at processing the text within a given input, they lack a deeper contextual understanding of the situation or the broader world. This underscores the need for users to critically evaluate LLM-generated content as well as the need to have a certain degree of knowledge as to how to include context-relevant information within the input of a given LLM prompt.

Another major concern are biases presented in LLM outputs, especially as real-life incidents have shown how these biases can lead to problematic outcomes, raising serious ethical questions. For example, Schramowski et al. (2022) discuss how language models like BERT retain human-like biases, specifically in moral norms and values. LLMs are trained on vast datasets that often reflect biases present in society (Abid et al., 2021), and users who are unaware or dismissive of these biases may inadvertently perpetuate or amplify them through uncritical use of LLM outputs. For example, if an LLM is used to screen job applications and the training data had biases against certain demographic groups, the LLM might (somewhat predictably) replicate these biases in its screening process, leading to unfair and discriminatory practices.[4]

It is important to stress that a significant portion of the risks associated with LLMs stem not just from the models themselves, but also from a lack of understanding about how to use them effectively and responsibly. For example, understanding the limits of the context windows of a particular LLM becomes highly relevant when using large inputs; one might risk believing the model has incorporated certain information (e.g., when summarizing a large document) when this is not the case. Often, pitfalls encountered in practical applications of LLMs can be traced back to basic level gaps in knowledge about certain system's limitations or to misconceptions about the capabilities and limitations of these technologies in general.

[2] Importantly, our concern is with mitigation of unintentional or possibly negligent misuse stemming from user ignorance regarding the limitations of these systems. Willful and malicious misuse will still obviously present a problem, but mitigation strategies for this will need to be crafted along very different lines, in keeping with the different nature of such misuses. Exploration of this is beyond the scope of the current article.

[3] See, e.g., (Augenstein et al., 2023; Barman et al., 2024a, 2024b; Chen & Shu, 2023; Mittelstadt et al., 2023).

[4] For examples of such problems arising in real-world contexts, see, e.g., Gallegos et al. (2023), Li et al. (2023) and Salinas et al. (2023).

To navigate these pitfalls, it is essential to adopt guidelines for responsible LLM usage, which include ethical considerations, bias mitigation strategies, and regular accuracy checks. Educating users on the capabilities and limitations of LLMs is also paramount, ensuring that they are leveraged as tools that augment human abilities rather than replace them. Training programs and educational resources can play a significant role in enhancing the understanding and effective utilization of LLMs.

The challenges of XAI for LLMs

Explainable Artificial Intelligence (XAI) has emerged as a significant field in the study of AI, aimed at making the decisions and processes of AI systems transparent and understandable to humans. Its relevance for LLMs is particularly noteworthy, considering the increasing reliance on these models in various domains and the significant degree of opacity in most LLMs. However, the task of elucidating the inner workings of LLMs through XAI presents unique challenges.

The fundamental complexity and scale of LLMs presents significant hurdles. These models, characterized by extremely large parameter counts (e.g., GPT4 is said to have 1.7 trillion parameters), process and generate information in ways that are not intuitively understandable. XAI explanations may not only be inadequate, or computationally costly (given the size of these models) but can potentially mislead users into a false sense of understanding. Moreover, as will be argued below, even if these explanations are good explanations of the 'internal reasoning' of the model, or of the reasons for a particular output being what it is, they might not convey the actionable information needed for adequate use.

XAI techniques are a host of computational, mathematical, and statistical models that, when applied to AI models, enable gathering key information about their workings. There are many approaches to be found, ranging from trying to figure out the importance certain features might have towards an output (Lundberg & Lee, 2017), to providing simplified linear models that track some level of accuracy in the behavior of the model within a smaller range of inputs (Ribeiro et al., 2016), to trying to tease out the mechanisms of neural network behavior (e.g., by highlighting the circuits, or by showing which input maximally activates a certain neuron), which is sometimes called mechanistic explainability (Conmy et al., 2023a, 2023b).

Some of the most promising XAI techniques for studying LLMs use other bigger LLMs to study key features of smaller ones (Bills et al., 2023), but it is fair to say that the current results are modest (see Zhao et al., 2023 for a recent summary on XAI for LLMs). Model-agnostic methods (Zolanvari et al., 2021) might be helpful to some extent when trying to provide explanations for given outputs (i.e., local explanations). The strength of these methods is that they can operate regardless of the model architecture. This sidesteps the need to delve into the intricate and often incomprehensible internal mechanisms of these complex models. For instance, examples such as SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), aim to provide insights into the decision-making process of AI models by approximating how changes in input affect the model's output. The main issue here is that knowing the relative importance of certain components of the input towards an output will likely not provide the necessary information to improve future user actions.

Supposing that one can adequately establish that certain nouns and phrases of the input to an LLM were very important to the output, this provides little information as to whether the output is correct, as to whether the output is biased or incomplete, or as to what should have been changed in the input so that the output would have been better. (It is worth mentioning that counterfactual explanations for LLMs are still in their infancy, although this could be a promising way of resolving some of these issues, especially when combined with interventionist approaches; see e.g., Meng et al. (2022). We are, however, skeptical that this would be enough to solve all issues mentioned so far.)

The broader question that arises is whether these explanations are the 'right' sort of thing for addressing the challenges posed by LLMs. While model-agnostic techniques can provide a snapshot of how an LLM might be processing and responding to specific inputs, they may fall short of conveying the broader context and inherent uncertainties of LLM functioning.

Implementing effective XAI solutions is not only technically challenging but also resource-intensive, often requiring additional computational power and specialized expertise. But even if these obstacles proved trivial, practical implementations of XAI with LLMs also encounter substantial issues. First of all, there is a risk of user misunderstanding, where the explanations provided by XAI tools might lead users to overestimate their understanding or make incorrect assumptions about the LLM's outputs (that is, assuming users can understand the explanations in the first place). Real-world examples further illustrate these challenges. For instance, an LLM used in a legal context might provide a decision rationale that seems plausible to a layperson but misses critical legal nuances, leading to misinterpretation of its advice. Conversely, there are scenarios where XAI successfully sheds light on certain aspects of an LLM's decision process but still leaves a gap in comprehensive understanding, particularly in cases involving nuanced or abstract concepts.

At a more fundamental level, XAI may prove less helpful than anticipated for improving responsible and reliable use of LLMs, for the simple reason that knowing *why* some outcome is reached may not aid particular users in getting the outcome they are seeking. This is because knowledge of *why* a system functions as it does sometimes will not translate to knowledge of *how* to actually get something done, and it is possible for users, especially more beginner users like students, to see explanations and not know what their implications are for LLM usage. Furthermore, due to the general audience being considered there are two potential risks. Firstly, the explanations might be misunderstood. Secondly, explanations might be oversimplifying in undesirable ways to the user's goal. What is critical is thus not that users have some explanation, however good it might be, but rather that they know the dos-and-don'ts of responsibly using that system. This is especially important for more novice users, for the simple fact that dos-and-don'ts can be effectively and reliably communicated to such individuals, while explanations (and what to do with those) may be unhelpful or may require an initial learning phase to get acquainted with said explanations.[5]

Given these various limitations, alternative approaches to understanding LLM outputs are worth further exploration. Human-centric approaches that focus on educating users about the general principles and potential biases of LLMs can be beneficial. Alongside XAI, complementary tools and methods, such as heuristic-based analysis, can provide a more rounded understanding of LLM outputs.[6]

The role of user training and the development of practical guidelines cannot be overstated in this context. To effectively harness the capabilities of LLMs, users may not need to fully understand AI systems themselves, but they will need a baseline competence in understanding both *how to use* AI systems and they will have to have a clear idea of the limitations of these systems. Practical guidelines that offer realistic and user-friendly advice on input crafting, and on interpreting and applying LLM outputs are crucial. These guidelines should be developed in tandem with advancements in XAI, ensuring they remain relevant and useful. Put more bluntly, if the goal is to enable users to utilize LLMs in a responsible, reliable, and efficient manner, simply providing explanations might not suffice, even if good explanations are available.

Charting a user-centric path: the case for tailored guidelines in a user-centric-approach (UCA)

Specific LLMs like GPT-4 are still in their early stages of development, yet already even experts do not fully comprehend the intricate details of how these models work or their full capabilities and limitations.[7] Expecting explainability from such nascent and complex systems often sets unrealistic expectations. While explainability is a worthy long-term goal, over-prioritizing it now directs attention and resources away from developing the core functionalities and safety protocols which should be the current prime focus.[8] This is not to say that both should not work in tandem in a complementary fashion, as breakthroughs in XAI will likely inform good policies and guidelines, but we need not wait for such breakthroughs to already start providing good use norms.

Crafting user-centric guidelines is a fundamental step towards a more practical and inclusive use of LLMs. Such guidelines should be dynamic, evolving in response to technological advancements and user feedback. This flexibility is particularly crucial in areas like healthcare, legal services, and education, where LLMs are increasingly being integrated. By catering to these diverse requirements, we can promote a more inclusive and responsible application of LLMs, reducing risks and maximizing their potential.

Additionally, developing these guidelines should be a collaborative endeavor, involving AI experts, ethicists, educators, and perhaps most importantly, end-users. This broad spectrum of perspectives can help to ensure that guidelines more comprehensively address everything from technical proficiency to ethical concerns. Utilizing real-world case studies of LLM applications, both successful and problematic, can provide invaluable insights. These examples can serve as a foundation for developing practical, context-specific strategies. For instance, guidelines could offer advice on identifying and mitigating biases in LLM outputs, a significant ethical concern.

For the average user, simple heuristics to detect potential errors or misinformation from LLMs may be far more beneficial than obtaining an explanation for a given output. Pattern recognition to identify peculiar phrasings, internal contradictions, or suspicious cited sources requires little technical skill. Equipping the public with easily comprehensible rules of thumb can empower users to exercise caution with LLM outputs and to use these tools effectively.

[5] See Wood (2024) for further exploration of challenges in using XAI to improve effective and responsible use of AI-enabled systems.

[6] See, e.g., Liao and Vaughan (2023) and Wang et al. (2024a, 2024b).

[7] E.g., Bowman (2023) and Zhao et al. (2023). See also the discussion presented in layman's terms at https://www.linkedin.com/pulse/when-llm-experts-say-we-dont-know-how-pallav-sharda-2tpyc/.

[8] More broadly, emphasis on XAI, assuming it can be fully achieved, may undermine more institutional and human-centric approaches. See Wood (2024).

Additionally, this approach places emphasis on the core points at stake, namely responsible and reliable use of LLMs, and does not allow for undue attention on technical achievements like explainability which may not always foster responsible and reliable use in all contexts.

Creation of structured protocols and best practices can further prevent unintentional misuse and mitigate risks even without explainability. Educational and regulatory institutions can (and arguably should) collaborate to develop appropriate regulations tailored to different disciplines, as detailed protocols with specific use cases, limitations, and oversight can constrain the risks of LLMs irrespective of explainability. These protocols and practices can be tested and regimented through benchmarks (see e.g., Barman et al., 2024a, 2024b).

In any case, LLMs are still maturing technologies and their risks can arguably be managed through usual protocols of appropriate regulation, education, and responsible development, as with any powerful new technology. The internal combustion engine is not easily explainable to average drivers, yet does not imperil them when proper precautions are instituted. Similarly, LLMs can be incorporated into society once protocols for safe usage are established. For typical users, explainability is not an essential prerequisite for benefitting from LLMs' advantages.

For ordinary users, having an explanation may not be necessary (and certainly not sufficient) to handle routine tasks. Just as a light switch user need not comprehend electrical engineering, appropriate LLM usage can be taught without elucidating the reasons, details, or inner workings of models. The essential knowledge required is an understanding of capabilities, limitations, and proper use, not of the underlying model itself.

Issues regarding appropriate development and regulation of AI are thus also institutional challenges rather than purely technical ones. Technical solutions like explainability, benchmarking, and error analysis will likely supplement broader initiatives like industry standards, responsible research norms, and educational curricula. Simple warning signals and guidelines could enable safe LLM usage for regular tasks without necessitating intricate explainability. The right level of transparency must align with the user and use case. But the onus lies on educational and regulatory institutions to spearhead research into challenges arising from real-world LLM usage, and accordingly, formulate policies and protocols to ensure these models are employed for the collective good.

Meta criteria for guidelines

To harness the full potential of LLMs while mitigating their inherent risks, the development and implementation of effective guidelines is a critical undertaking. The aim of such guidelines ought to be to minimize risks while maximizing the potential of LLMs. As such, these guidelines should also either minimize the number of errors or misinformation, or minimize the severity of mistakes. On the other hand, these guidelines should also be specific enough to provide contextual heuristics that would allow users to craft prompts that are satisfactory.

In the following, we will not propose any definitive guidelines, but instead propose desiderata or meta criteria that such guidelines should satisfy (and explore possible guidelines which might follow from these). Adherence to these meta criteria should ideally ensure that LLM's use is practical, reliable, and beneficial. Central to these criteria is the notion that guidelines should be empirically testable,[9] tailored to specific audiences and use cases, and integrated into the very interfaces of LLMs, much like the disclaimers provided by tools like ChatGPT about potential inaccuracies in information. However, we believe additional institutional aid is required on this matter.

First and foremost, guidelines must be evidence-based. They should be testable and tested to validate their effectiveness. This evidence-based approach ensures that guidelines are more than just theoretical constructs; they are practical tools that have been proven to work in real-world scenarios. In this context, efficiency is a key consideration. The guidelines should streamline users' interaction with LLMs, improving the overall experience without adding unnecessary complexity or cognitive load.[10] They should also show promising results when it comes to minimizing potential costs or losses due to misuse or mistakes, be these costs economic, reputational, social, health-related or of another nature. In this sense, it is important that guidelines not only

[9] Some might argue that "rules of thumb" or heuristics for guiding LLM use are not apt to empirical testing or verification. What we have in mind, however, is a general ability to empirically check whether guidelines improve use of LLMs (in terms of users accomplishing the tasks they are employing LLMs for), and in this respect, it should be possible to empirically examine whether guidelines are indeed improving use, detracting from it, or having a negligible impact. The precise impact of various guidelines, and their implementation, would further provide useful running data for the improvement of user interfaces with an eye to ever more effective and responsible LLM use. See also Barman et al., (2024a, 2024b).

[10] For candidate approaches in this direction, see, e.g., Wang et al. (2024a, 2024b and Watkins (2023) as well as https://www.dpc.sa.gov.au/__data/assets/pdf_file/0007/936745/Guideline-13.1-Use-of-Large-Language-Model-AI-Tools-Utilities.pdf and https://www.isc.upenn.edu/security/LLM-guide. See Johri et al. (2023) for more meta-level guidelines embedded within a specific context, i.e., LLM use in the field of medicine.

minimize the number of mistakes, but also the magnitude of these mistakes.

It is also important to acknowledge that the potential use cases where LLMs are limited or prone to errors do not necessarily preclude their general or overall utility. Guidelines should be directed at specific users employing LLMs for specific purposes, and in some cases, the occasional inaccuracies of LLMs might be acceptable, provided the user is aware of these limitations and responds accordingly.

The role of institutions in providing these guidelines is also a point of consideration. Whether its academic bodies, industry leaders, or regulatory agencies, the responsibility for developing and disseminating these guidelines might fall upon various stakeholders. Regardless of who ultimately takes on this task, the emphasis should be on collaborative efforts that draw on diverse expertise and perspectives.

This becomes an especially important issue considering the fact that lack of proper guidelines disproportionately affects lower socio-economic groups who might lack resources or meta-skills. These groups are often the most vulnerable to negative consequences of LLM errors or misuse due to limited access to alternative information sources or support systems. Inadequate guidelines could result in these users being more susceptible to misinformation, digital manipulation, or opportunity costs. Therefore, it's crucial that guidelines not only cater to the technological aspects of LLMs, but also consider socio-economic disparities. This involves creating accessible, easy-to-understand guidelines that can be effectively utilized by individuals from various backgrounds. It also means ensuring that the guidelines are distributed widely and are available in multiple languages and formats, to reach a broader audience. Such inclusive approaches in guideline development and dissemination are vital in ensuring that the benefits of LLMs are equitably shared and the risks are minimized across all societal strata.

However, the creation of guidelines is just the beginning. Their real-world efficacy must be continuously tested and refined. This iterative process of testing and refinement is essential to ensure that guidelines remain relevant and effective as LLMs evolve and new challenges emerge.

In educational settings, the incorporation of LLMs must be guided by well-established best practices to make the learning experience more personalized and efficient. These practices should vary according to the level of education. For instance, the way an elementary school student engages with an LLM should differ significantly from a graduate student's approach. It is crucial that students at all levels learn not only to use these tools but also to critically analyze their outputs.

Educators will have a central role in this process, but also the institutions related to education. These should provide guidance on the ethical and responsible use of LLMs, tailored to various tasks and contexts. This involves teaching students not just how to use these models but also how to verify and attribute the information provided by them correctly. This is thus akin to learning citation principles; students must understand how to acknowledge and evaluate LLM contributions accurately. Furthermore, guidelines should include good principles for how to make best use of these models, taking advantage of their strengths. A critical aspect of this training involves providing students with adequate instruction in prompt engineering principles and core strategies (see e.g., Buruk, 2023). This could include techniques like chain of thought prompting (Wei et al., 2022) or assigning specific roles to the model to guide its responses. Students should also be instructed in how to evaluate the validity and reliability of the outputs they receive. This involves fact-checking strategies and developing an understanding of the strengths and limitations of LLMs.

Moreover, it is essential to establish good heuristics for prompt iteration and to create a work system that integrates these tools seamlessly into the students' workflow. This integration should aim at enhancing the educational process, encouraging critical thinking, and fostering a deeper understanding of how to interact with advanced AI tools like LLMs effectively.

When it comes to dealing with failures or misuses of LLMs, these instances shouldn't be viewed merely as mistakes, but rather as valuable learning opportunities. Each failure provides a unique chance to revisit and refine the training protocols and curricula surrounding the use of these technologies. Educators can use these experiences to highlight the importance of responsible LLM use, demonstrating the consequences of misuse and the steps needed to correct or avoid such situations in the future. In addition, such experiences should be integrated into educational curricula as case studies or examples, allowing students to learn from real-life scenarios. This approach not only makes the learning process more relatable but also prepares students for the practical challenges they may face in using AI tools in their future academic or professional pursuits.

In workplace settings, similar principles are essential. Initially, well-defined guidelines can significantly enhance workforce productivity and it has already been demonstrated that the use of LLMs positively affects productivity (Eloundou et al., 2023). Enhancing these effects further can be achieved through a robust set of guidelines for effective prompting. In the case of legal settings, clear guidelines help delineate the type of information that can be legally shared with LLMs, ensuring ethical compliance and the protection of sensitive data. These guidelines are particularly crucial for employees who are less technologically savvy and might otherwise hesitate to utilize LLMs in their professional roles.

When it comes to expert advice, such as seeking legal or health guidance from LLMs, the guidelines should primarily aim at mitigating the potential impact of inaccuracies.

This involves emphasizing the importance of treating the information provided as a starting point rather than a definitive solution. Users should be encouraged to verify critical details with qualified professionals. Additionally, the guidelines should highlight the importance of discretion when dealing with sensitive or personal issues, and clearly state the limitations of LLMs in understanding and interpreting complex, nuanced situations. By doing so, the guidelines can help to maintain a realistic expectation of the assistance LLMs can provide, while underscoring the importance of professional judgment and expertise in these specialized areas.

Conclusion

Large language models are having a significant impact across diverse sectors. However, these systems provide not just possibilities for benefit, but also for error, misunderstanding, and even harm. In order to address these risks, a nuanced and responsible approach centered on user education and training is needed. More than this, mitigating the risks associated with LLMs, such as biases or misinformation, requires more than just an emphasis on explainability. It demands a concerted effort towards developing robust user guidelines, gathering insights from actual LLM usage, and addressing challenges encountered in real-world applications.

Central to our discussion is the need for thoughtfully designed curricula and basic training programs and user guidelines that cater to different users' varying experience levels and disciplines. These programs should crucially focus on ethical and responsible usage. Education on prompt engineering and guidelines for differentiating appropriate from inappropriate LLM tasks is imperative. This approach further aligns with the need for a sociotechnical perspective that emphasizes human guidance and oversight in the use of LLMs.

Our exploration reveals that realizing the full potential of LLMs extends beyond the technical realm, and in fact may most heavily demand an institutional rather than technical approach to minimizing error and misuse. It calls for a holistic strategy that integrates prudent oversight and education. Such a strategy ensures that these powerful models serve as a force for positive transformation in various aspects of society. By placing a premium on the human element in the integration of LLMs, we can pave the way for an ethical and responsible framework that maximizes the benefits of these advanced technologies across institutions.

Future research should focus on the empirical evaluation of the effectiveness of user training programs and guidelines in enhancing the safe and ethical use of LLMs. This involves conducting systematic studies across various contexts to understand how different user groups interact with LLMs and the impact of specific training protocols on their ability to use these models responsibly. Further work could also explore the development of adaptive training systems that evolve based on user feedback and the changing dynamics of LLM technologies. Additionally, there is a need for interdisciplinary studies that integrate insights from fields such as ethics, psychology, and education to further refine and contextualize user guidelines. Another promising area involves the investigation of the long-term social and ethical impacts of LLM use in different sectors, such as education, healthcare, and business, to inform the continuous refinement of user guidelines. This research should aim to establish a comprehensive framework that not only guides users in the present but also anticipates future challenges and opportunities in the realm of artificial intelligence.

Funding This work was funded by Fonds Wetenschappelijk Onderzoek (Grant numbers: 1229124N for Kristian González Barman and 1255724N for Pawel Pawlowski) and the Czech Science Foundation (Grant number 24-12638I for Nathan Wood).

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. *Nature Machine Intelligence, 3*(6), 461–463.

Agarwal, V., Thureja, N., Garg, M. K., Dharmavaram, S., & Kumar, D. (2024). "Which LLM should I use?": Evaluating LLMs for tasks performed by Undergraduate Computer Science Students in India. Preprint retrieved from arXiv:2402.01687.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82–115.

Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G. L., Corney, D., ... & Zagni, G. (2023). Factuality challenges in the era of large language models. Preprint retrieved from arXiv:2310.05189.

Barman, D., Guo, Z., & Conlan, O. (2024). The dark side of language models: Exploring the potential of LLMs in multimedia disinformation generation and dissemination. *Machine Learning with Applications, 16*, 100545.

Barman, K. G., Caron, S., Claassen, T., & De Regt, H. (2024b). Towards a benchmark for scientific understanding in humans and machines. *Minds and Machines, 34*(1), 1–16.

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

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., & Saunders, W. (2023) Language

models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.Html
Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines, 32*(1), 43–75.
Boiko, D. A., MacKnight, R., & Gomes, G. (2023). Emergent autonomous scientific research capabilities of large language models. Preprint retrieved from https://arxiv.org/abs/2304.05332
Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 2053951715622512.
Buruk, Oğuz'Oz. (2023) Academic Writing with GPT-3.5: Reflections on practices, efficacy and transparency. Preprint retrieved from arXiv:2304.11079.
Chen, C., & Shu, K. (2023). Combating misinformation in the age of LLMs: Opportunities and challenges. Preprint retrieved from arXiv:2311.05656.
Choi, E. (2023). A comprehensive inquiry into the use of ChatGPT: Examining general, educational, and disability-focused perspectives. *International Journal of Arts Humanities and Social Sciences*. https://doi.org/10.56734/ijahss.v4n11a1
Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. Preprint retrieved from arXiv:2304.14997.
Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023b). Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems, 36*, 16318–16352.
de Fine Licht, K. (2023). Integrating large language models into higher education: guidelines for effective implementation. *Computer Sciences & Mathematics Forum, 8*(1), 65.
Dergaa, I., Chamari, K., Zmijewski, P., & Ben Saad, H. (2023). From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport, 40*(2), 615–622. https://doi.org/10.5114/biolsport.2023.125623
Durán, J. M. (2021). Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. *Artificial Intelligence, 297*, 103498.
Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. Preprint retrieved from arXiv:2303.10130.
Essel, H. B., Vlachopoulos, D., Essuman, A. B., & Amankwa, J. O. (2024). ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). *Computers and Education: Artificial Intelligence, 6*, 100198.
Extance, A. (2023). ChatGPT has entered the classroom: How LLMs could transform education. *Nature, 623*(7987), 474–477.
Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., & Hemphill, L. (2023). A bibliometric review of large language models research from 2017 to 2023. Preprint retrieved from https://doi.org/10.48550/arXiv.2304.02020
Fear, K., & Gleber, C. (2023). Shaping the future of older adult care: ChatGPT, advanced AI, and the transformation of clinical practice. *JMIR Aging, 6*(1), e51776.
Ferrara, E. (2023). Should chatgpt be biased? Challenges and risks of bias in large language models. Preprint retrieved from arXiv:2304.03738.
Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2023). Bias and fairness in large language models: A survey. Preprint retrieved from arXiv:2309.00770.
Girotra, K., Meincke, L., Terwiesch, C., & Ulrich, K. T. (2023). Ideas are dimes a dozen: Large language models for idea generation in innovation. Available at SSRN 4526071.
Guo, Y., & Lee, D. (2023). Leveraging chatgpt for enhancing critical thinking skills. *Journal of Chemical Education, 100*(12), 4876–4883.
Hadi, M. U., Al-Tashi, Q., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., Mirjalili, S., & Shah, M. (2023). Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. Preprint retrieved from https://doi.org/10.36227/techrxiv.23589741.v4
Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. Authorea Preprints.
Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169*, 615–626.
Inagaki, T., Kato, A., Takahashi, K., Ozaki, H., & Kanda, G. N. (2023). LLMs can generate robotic scripts from goal-oriented instructions in biological laboratory automation. Preprint retrieved from https://doi.org/10.48550/arXiv.2304.10267
Jablonka, K. M., Ai, Q., Al-Feghali, A., Badhwar, S., Bocarsly, J. D., Bran, A. M., Bringuier, S., Brinson, L. C., Choudhary, K., Circi, D., Cox, S., de Jong, W. A., Evans, M. L., Gastellu, N., Genzling, J., Gil, M. V., Gupta, A. K., Hong, Z., Imran, A., ... Blaiszik, B. (2023). 14 examples of how LLMs can transform materials science and chemistry: A reflection on a large language model hackathon. *Digital Discovery, 2*(5), 1233–1250. https://doi.org/10.1039/d3dd00113j
Johri, S., Jeong, J., Tran, B. A., Schlessinger, D. I., Wongvibulsin, S., Cai, Z. R., ... & Rajpurkar, P. (2023). Guidelines for rigorous evaluation of clinical LLMs for conversational reasoning. medRxiv, 2023–09.
Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, *103*, 102274.
Kim, J. K., Chua, M., Rickard, M., & Lorenzo, A. (2023). ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology., 19*, 598.
Lee, J., Le, T., Chen, J., & Lee, D. (2023). Do language models plagiarize? *In Proceedings of the ACM Web Conference 2023* (pp. 3637–3647). ACM. https://doi.org/10.1145/3543507.3583199
Li, Y., Du, M., Song, R., Wang, X., & Wang, Y. (2023). A survey on fairness in large language models. Preprint retrieved from arXiv:2308.10149.
Liao, Q. V., & Vaughan, J. W. (2023). Ai transparency in the age of llms: A human-centered research roadmap. Preprint retrieved from arXiv:2306.01941
Lin, Z. (2023). Why and how to embrace AI such as ChatGPT in your academic life. *Royal Society Open Science, 10*(8), 230658. https://doi.org/10.1098/rsos.230658
Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30.*
Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems, 35*, 17359–17372.
Mishra, A., Soni, U., Arunkumar, A., Huang, J., Kwon, B. C., & Bryan, C. (2023). Promptaid: Prompt exploration, perturbation, testing and iteration using visual analytics for large language models. Preprint retrieved from arXiv:2304.01964.
Mittelstadt, B., Wachter, S., & Russell, C. (2023). To protect science, we must use LLMs as zero-shot translators. *Nature Human Behaviour, 7*(11), 1830–1832.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Available at SSRN 4375283*.

OpenAI, R. (2023). Gpt-4 technical report. Preprint retrieved from arxiv:2303.08774. View in Article, 2.

Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.-Y., & Wang, W. Y. (2023). On the risk of misinformation pollution with large language models. Preprint retrieved from https://doi.org/10.48550/arXiv.2305.13661

Qadir, Junaid. (2023) Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. *2023 IEEE Global Engineering Education Conference* (*EDUCON*). *IEEE, 2023*.

Rakap, S. (2023). Chatting with GPT: Enhancing individualized education program goal development for novice special education teachers. *Journal of Special Education Technology*. https://doi.org/10.1177/01626434231211295

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. Preprint retrieved from arXiv:1606.05386.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Salinas, A., Shah, P., Huang, Y., McCormack, R., & Morstatter, F. (2023, October). The Unequal Opportunities of Large Language Models: Examining Demographic Biases in Job Recommendations by ChatGPT and LLaMA. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1–15).

Schramowski, P., Turan, C., Andersen, N., & Herbert, F. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence, 4*(3), 258–268. https://doi.org/10.1038/s42256-022-00458-8

De Silva, D., Mills, N., El-Ayoubi, M., Manic, M., & Alahakoon, D. (2023). ChatGPT and generative AI guidelines for addressing academic integrity and augmenting pre-existing chatbots. *In 2023 IEEE International Conference on Industrial Technology* (*ICIT*) (pp. 1–6). IEEE. https://doi.org/10.1109/ICIT58465.2023.10143123

Sun, Z. (2023). A short survey of viewing large language models in legal aspect. Preprint retrieved from arXiv:2303.09136.

Valentino, M., & Freitas, A. (2022). Scientific explanation and natural language: A unified epistemological-linguistic perspective for explainable AI. Preprint retrieved from arXiv:2205.01809.

Vidgof, M., Bachhofner, S., & Mendling, J. (2023). Large language models for business process management: Opportunities and challenges. Preprint retrieved from https://doi.org/10.48550/arXiv.2304.04309

Wang, J., Ma, W., Sun, P., Zhang, M., & Nie, J. Y. (2024). Understanding user experience in large language model interactions. Preprint retrieved from arXiv:2401.08329.

Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., & Li, J. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine, 7*(1), 41.

Watkins, R. (2023). Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00294-5

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903

Williams, N., Ivanov, S., & Buhalis, D. (2023). Algorithmic ghost in the research shell: Large language models and academic knowledge creation in management research. Preprint retrieved from https://doi.org/10.48550/arXiv.2303.07304

Wood, N. G. (2024). Explainable AI in the military domain. *Ethics and Information Technology, 26*(2), 1–13.

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 75–78). ACM. https://doi.org/10.1145/3581754.3584101

Yadav, G. (2023). Scaling evidence-based instructional design expertise through large language models. Preprint retrieved from https://doi.org/10.48550/arXiv.2306.01006

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic literature review. Preprint retrieved from https://doi.org/10.48550/arXiv.2303.13379

Yell, M. M. (2023). Social studies, ChatGPT, and lateral reading. *Social Education, 87*(3), 138–141.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... & Du, M. (2023). Explainability for large language models: A survey. Preprint retrieved from arXiv:2309.01029.

Zolanvari, M., Yang, Z., Khan, K., Jain, R., & Meskin, N. (2021). Trust xai: Model-agnostic explanations for ai with a case study on iiot security. *IEEE Internet of Things Journal.*