**Discover** Artificial Intelligence

Perspective

# LLM potentiality and awareness: a position paper from the perspective of trustworthy and responsible AI modeling

**Iqbal H. Sarker**[1,2]

Received: 1 March 2024 / Accepted: 8 May 2024
Published online: 21 May 2024

**Abstract**
Large language models (LLMs) are an exciting breakthrough in the rapidly growing field of artificial intelligence (AI), offering unparalleled potential in a variety of application domains such as finance, business, healthcare, cybersecurity, and so on. However, concerns regarding their trustworthiness and ethical implications have become increasingly prominent as these models are considered black-box and continue to progress. This position paper explores the potentiality of LLM from diverse perspectives as well as the associated risk factors with awareness. Towards this, we highlight not only the technical challenges but also the ethical implications and societal impacts associated with LLM deployment emphasizing fairness, transparency, explainability, trust and accountability. We conclude this paper by summarizing potential research scopes with direction. Overall, the purpose of this position paper is to contribute to the ongoing discussion of LLM potentiality and awareness from the perspective of trustworthiness and responsibility in AI.

**Keywords** LLM · Automation · Risk factors · Trustworthy AI · Responsible AI · Text analytics · NLP · Machine learning · Data science

## 1 Introduction

Large language models (LLMs) have been widely used in recent years, revolutionizing the area of artificial intelligence (AI) by allowing machines to process and produce human-like text on an unprecedented scale. These models such as OpenAI's GPT (Generative Pre-trained Transformer) series and Google's BERT (Bidirectional Encoder Representations from Transformers) [1], leveraging deep learning architectures and massive datasets, have demonstrated remarkable capabilities in understanding, generating, and manipulating human-like text.

LLMs have facilitated the development of innovative applications, ranging from chatbots and virtual assistants to content creation tools and personalized recommendation systems, discussed briefly in Sect. 4. However, alongside their transformative potential, LLMs also pose significant challenges and risks that necessitate careful consideration. One significant concern associated with LLM is its critical tendency to produce hallucinations, i.e., generating content that appears nonsensical or unfaithful [2]. Algorithmic bias is another concern, which eventually leads to unfair or discriminatory outcomes. These tools can occasionally produce extremely convincing but possibly false information, which raises questions about the dissemination of misinformation and disinformation. Moreover, the black-box nature of LLMs poses challenges in understanding how and why certain outputs are generated [3]. Furthermore, adversarial attacks, in which malicious actors take advantage of vulnerabilities to manipulate model outputs or undermine system integrity, can

✉ Iqbal H. Sarker, m.sarker@ecu.edu.au | [1]Centre for Securing Digital Futures, Edith Cowan University, Perth, WA 6027, Australia. [2]Cyber Security Cooperative Research Centre, Perth, WA 6027, Australia.

target LLMs due to their immense size and complexity. Addressing these risk factors needs to be taken into account in LLM modeling practices, which are discussed briefly in Sect. 5.

This paper seeks to explore mainly trustworthy and responsible AI modeling, aiming to provide insights into the ethical considerations and challenges associated with the development and deployment of LLM systems. While "trustworthy AI" focuses on technical reliability, transparency, and accountability, "responsible AI" broadens the scope to encompass ethical considerations, societal impacts, and human values [3]. Together, these concepts establish a framework for guiding the development, deployment, and governance of AI systems in a manner that fosters trust, equity, and societal well-being. To gain a deeper understanding of the core theme of this position paper and its overall contributions, we articulate three fundamental questions below:

- *LLM potentiality:* Can LLM effectively process and analyze massive amounts of data to develop real-world applications, which poses challenges for human analysts?
- *LLM risk factors and awareness:* What are the risk factors associated with LLMs, particularly concerning how we can ensure trustworthiness and responsibility?
- *Potential research scope:* What are the potential areas of research that can advance LLM-based solutions?

Overall, this position paper emphasizes the importance of comprehending the potential and awareness of LLM-based solutions within the context of trustworthy and responsible AI modeling. By highlighting the capabilities, challenges, and ethical dimensions, the paper aims to contribute to the establishment of best practices for designing, deploying, and governing LLM systems.

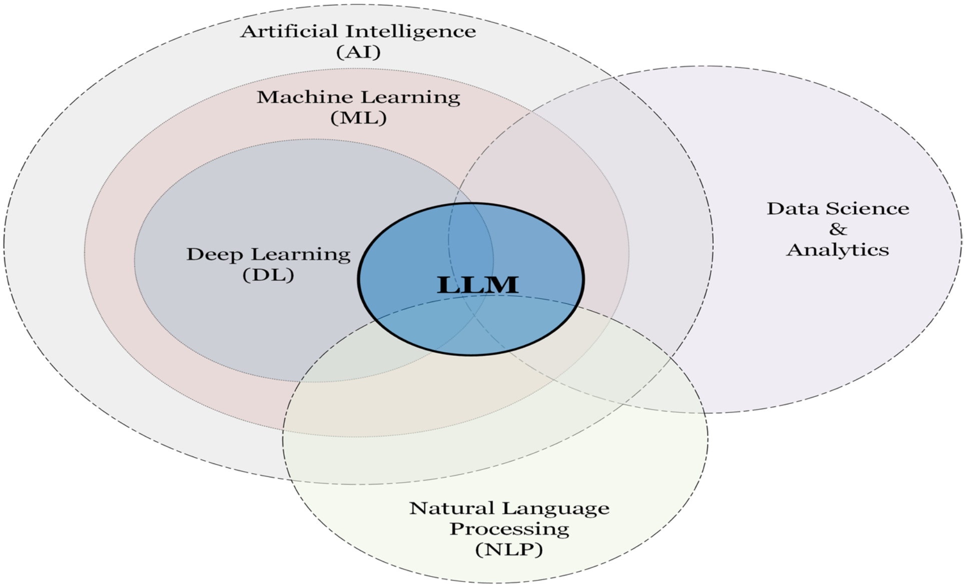
## 2 Understanding and positioning LLM

LLMs hold a significant position within the broader landscape of AI, data science, and natural language processing (NLP) research, reflecting their interdisciplinary nature and wide-ranging impact. Figure 1 illustrates the relevance and position of LLM within the area of AI, Data Science and NLP.

In the realm of AI, LLMs represent a significant milestone, showcasing the capabilities of machine learning (ML) and deep learning (DL) techniques [3] to understand and generate human-like text at an unprecedented scale. More specifically, their development and deployment rely heavily on neural networks, attention mechanisms, and transfer learning, highlighting the interconnectedness between LLMs and learning technologies within the broad area of AI.

Similarly, Data Science [4] serves as the backbone of LLM development, emphasizing the significance of advanced techniques in data analyzing and processing, ensuring data quality, and modeling techniques for achieving optimal performance, where expertise in data science is the key. Consequently, the knowledge of data science is pivotal in curating

**Fig. 1** An illustration highlighting the relevance and position of LLM within the area of AI, Data Science and NLP in a broader sense



and pre-processing datasets, fine-tuning and optimizing model architectures, and evaluating model performance, all of which are essential for the success of LLMs. Furthermore, LLMs contribute to advancing data science methodologies by introducing innovative techniques for handling unstructured text data, facilitating more sophisticated analysis and interpretation of textual information.

In the context of NLP research, LLMs serve as both a benchmark and a driving force, inspiring further innovation and exploration in areas such as language understanding, text generation, summarization, dialogue systems, machine translation and so on. Thus, LLMs drive research in NLP by offering new avenues for understanding language structure, semantics, and context, pushing the boundaries of language understanding and generation tasks. Their success has spurred the development of more sophisticated language models and techniques, fueling progress in the area [3].

In summary, the interaction between LLMs and the fields of AI, data science, and NLP research is mutually beneficial, encouraging interdisciplinary cooperation and advancing our comprehension and manipulation of human language. LLMs demonstrate the capabilities of AI systems in understanding and producing human language, emphasizing the significance of data science in model development and the contribution of NLP research in fostering innovation and tackling crucial challenges in language processing and comprehension. Their incorporation across these domains highlights the transformative influence of LLMs on AI, data science, and NLP research, emphasizing their pivotal role in shaping the future landscape of language-centric technologies and applications.

## 3 Key steps in LLM-based modeling

In this section, we briefly discuss the key modeling steps of LLMs ranging from problem definition to model deployment. These are:

- *Problem definition and task identification:* In this initial stage of LLM modeling, the main focus is on precisely defining the problem to be tackled. This involves articulating the nature of the input data, specifying the desired output, and clarifying the objectives of the task, whether it be classification, generation, translation, or other tasks. By carefully defining the problem statement and task requirements, researchers can efficiently use LLMs to address a wide range of real-world challenges across various domains and applications
- *Data acquisition and preprocessing:* The primary emphasis at this stage is on obtaining high-quality data and readying it for model training. This includes sourcing pertinent datasets that correspond to the task objectives, ensuring data quality, and diversity, and mitigating potential biases. Several data preprocessing techniques [4] can be employed to cleanse, standardize, and reshape the data into a format suitable for LLM training. By carefully selecting and preprocessing the data, researchers lay the foundation for building precise and trustworthy LLMs that can accurately represent the nuances of natural language.
- *Model selection and fine-tuning*: During this stage of LLM modeling, the key objective is to pinpoint the most fitting pre-trained LLM architecture and refine it for the specific task at hand. This phase involves assessing various pre-trained LLM models, such as GPT, BERT, or RoBERTa, considering factors like model size, computational resources, and task requirements [1]. Once an appropriate pre-trained model is chosen, fine-tuning techniques are applied to adjust the model parameters to suit the target task. This typically involves training the model on task-specific data while updating only a subset of the model's parameters, leveraging transfer learning to facilitate convergence and enhance performance. Through meticulous model selection and fine-tuning, researchers can customize LLMs to effectively tackle a broad spectrum of real-world tasks, achieving better accuracy and efficiency.
- *Model evaluation:* At this stage, the main objective is to evaluate the performance and efficacy of the trained model comprehensively. This involves rigorous testing and assessment utilizing validation datasets and customized metrics specific to the task or application. Researchers utilize a range of evaluation metrics, including accuracy, precision, recall, F1-score [3], and others according to the task type, to assess the model's performance across various dimensions. Furthermore, qualitative analyses such as human evaluation or error analysis may be employed to gain insights into the model's strengths and weaknesses. The overall goal is to ensure that the LLM satisfies the desired criteria for accuracy, generalization, and robustness, validating its suitability for deployment in real-world scenarios.
- *Deployment and monitoring:* This step involves deploying the fine-tuned LLM in a production environment or integrating it into an application pipeline for real-world usage. Monitoring the model's performance, detecting drift, and gathering user feedback to maintain the ongoing reliability and effectiveness in production, are all included.

By taking into account these key steps, researchers and practitioners can effectively utilize the capabilities of LLMs for various tasks while ensuring robustness, interpretability, and reliability in model development and deployment. The key research scopes are summarized in Sect. 6.

## 4 LLM potentiality and usage scope

The adaptability of LLMs extends to various real-world applications, fundamentally altering how we engage with language-related tasks and transforming industries worldwide. One prominent area where LLMs demonstrate their potential is in natural language understanding and generation [1]. In the following, we outline several application domains, where LLMs can make substantial contributions:

- *Financial analysis and decision-making:* LLMs have opened new possibilities for AI applications in finance [5]. By analyzing financial reports, market data, and economic trends, LLMs can assist with risk management, financial forecasting, and investment decisions. Their adeptness in interpreting intricate financial data, identifying anomalies, and developing forecasting models facilitates the optimization of portfolios and strategic planning.
- *News and social media analysis:* LLMs could be useful tools for identifying trends, sentiments, and misleading data in news stories, social media posts, and other online content. Thus these tools can support tasks like sentiment analysis, topic tracking, content verification, and so on.
- *Medical and healthcare:* This is one of the potential sectors of using LLMs [6]. LLMs can be used for tasks such as clinical documentation, medical coding, and patient communication. To gain important insights and assist in decision-making, they can examine a sizable volume of medical texts, including clinical notes, electronic health records, and medical literature. These tools can also help with diagnosis and treatment planning by supporting medical professionals in the synthesis and interpretation of complicated medical data. By analyzing and synthesizing medical literature, observing trends, and promoting knowledge discovery, they also play an essential role in medical research.
- *Education and e-learning:* LLMs can play a supportive role by offering automated tutoring, creating study materials, and enabling interactive learning experiences. These tools can aid in tasks such as information retrieval, concept comprehension, and language acquisition, thereby improving the efficiency and accessibility of education.
- *Cybersecurity:* This is another important sector where LLMs can play a key role [11, 12]. By analyzing a large volume of textual content, such as security reports, forums, and social media posts, LLMs can improve threat intelligence by identifying potential vulnerabilities and online risks [3]. Their ability to comprehend natural language allows them to identify phishing attempts, malware, and social engineering attacks more effectively. By offering context-rich insights into security events and suggesting appropriate actions, LLMs can also help automate incident response and threat remediation.
- *Customer service and virtual assistants:* LLMs can be implemented in chatbots for customer support to provide prompt answers to questions, troubleshoot problems, and assist users with product or service-related inquiries. By providing prompt and customized assistance, these models improve customer satisfaction while lightening the workload for human agents.

To sum up, large language models have the potential to revolutionize a wide range of real-world applications, increase productivity, and help people and organizations achieve their goals more precisely and swiftly. It is anticipated that as research and development into AI, machine learning, NLP, and data science advances, LLMs could grow increasingly important and influential in a variety of application areas in diverse sectors.

## 5 LLM risk factors and awareness

To ensure the responsible development and implementation of these powerful AI systems, it is important to comprehend the risks associated with LLMs and to increase awareness of these risks. In the following, we summarize some important LLM risk factors and areas of awareness:

- *Bias and fairness:* LLMs may unintentionally propagate biases found in the training data, resulting in unfair or discriminatory outcomes [7]. Increasing awareness involves identifying the origins of bias within the data and implementing strategies to mitigate bias throughout the model's development and deployment processes.

- *Misinformation and disinformation:* A significant challenge in the digital age is the spread of fake news [8]. LLMs may generate misleading or false information, posing risks to information integrity and trust. Thus awareness involves an in-depth understanding of the limitations of LLMs in assessing factual accuracy and establishing strategies to combat misinformation and disinformation.
- *Security vulnerabilities:* LLMs are susceptible to adversarial attacks, where malicious inputs may manipulate model outputs or compromise system security. Thus, awareness involves understanding and identifying the existence of security vulnerabilities in LLMs and deploying strong defenses, such as adversarial training and input validation, to safeguard against such attacks.
- *Privacy concerns:* LLMs trained on extensive datasets run the risk of unintentionally capturing sensitive or personally identifiable information, prompting privacy concerns. Raising awareness should emphasize the value of data privacy and promote the use of strong data anonymization and protection techniques when LLMs are being trained and deployed.
- *Algorithmic transparency and accountability:* LLMs are often considered as "black-box" models, making it challenging to understand their text generation or decision-making processes [3]. A greater understanding and trust in AI systems can be promoted by bringing attention to the difficulties of LLM transparency and establishing initiatives to create explainable AI methods and accountability mechanisms for LLM decisions.
- *Long-term effects:* Continuous monitoring and assessment are required because LLMs may have unexpected long-term effects on the economy, culture, and society. By conducting ongoing research and communication, awareness involves anticipating possible long-term effects of LLM deployment and proactively addressing emerging issues.
- *Ethical use:* LLMs raise ethical questions about where they should be used and where not and the possible effects on individuals or society. Creating a culture of ethical consciousness and responsibility among developers and users, outlining regulations for the responsible use of AI, as well as holding discussions regarding the ethical implications of LLM deployment are all part of raising awareness.
- *Education and literacy:* It is essential to enhance awareness of LLMs among the general people, policymakers, and stakeholders. Education initiatives can play a significant role in enabling individuals or organizations to comprehend the capabilities, limitations, and potential risks associated with LLMs, thereby empowering them to make informed decisions regarding their utilization and deployment.

Through comprehension of these risk factors and dissemination of information regarding them, consumers can take steps to optimize the benefits of LLMs while reducing their unfavorable impact on individuals, communities, and the broader society.

## 6 Key research scopes and directions

In general, the effectiveness and societal impact of AI/LLM technologies are shaped by a number of key issues (as pillars) including accuracy, explainability, trust, and ethical considerations [3, 9]. For instance, "accuracy" ensures that the models produce reliable outcomes, while "explainability" facilitates human understanding of their decision-making processes. "trust" is essential for user acceptance and adoption, requiring transparent and accountable modeling practices, particularly for real-world application areas. Additionally, "ethical considerations" encompass fairness, privacy, and societal impact, guiding the development of responsible AI systems. Prioritizing these key factors is crucial for developing and deploying trustworthy and responsible LLM systems that benefit individuals or society while minimizing the associated risks. To clearly understand these issues and finding potential research scope, we divide the LLM modeling phase into three categories—"Pre-modeling", "In-modeling" and "Post-modeling" phases, motivated from our earlier book by Sarker et al. [3]. In the following, we discuss each phase highlighting potential research scopes and directions, which might be helpful for the researchers as well as practitioners.

- *Pre-modeling phase:* In this pre-modeling phase of LLM, the research scope typically involves comprehensive data analysis and preparation to ensure fair, ethical, and unbiased model development from the perspective of trustworthy and responsible AI systems. This includes identifying potential sources of bias in the training data for a particular application and implementing strategies for bias mitigation and fairness-aware techniques into practice. To address biases and advance inclusivity, researchers need to set a high priority on gathering representative and diverse datasets, performing extensive data assessments, and employing careful preprocessing techniques. Data cleaning, mitigating

bias, augmentation, annotation, handling data poisoning issues or outliers, imbalance issues, and relevant others can be involved in real-world scenarios within the broad area of data science [4, 10]. Designing innovative algorithms or improving existing techniques to handle these data-specific issues and automation could be a novel contribution in this phase, depending on the data characteristics and nature. In addition, efforts should be focused on establishing clear objectives and ethical guidelines for the model's intended use in real-world application areas such as critical infrastructures, healthcare, business, cybersecurity, etc., as well as identifying any potential risk factors and ethical implications related to it.
- *In-modeling phase:* Research scope in this in-modeling phase typically emphasizes developing LLM architectures [1] or refining through parameter adjustment and training approaches that prioritize not only accuracy but also interpretability, fairness, and transparency. This involves investigating model architectures and training processes that enhance explainability and the ability of humans to interpret model results. It also involves developing models with fairness-aware algorithms and bias mitigation techniques. To further ensure the pre-trained LLM's accuracy and relevance in real-world applications, fine-tuning techniques are used to adapt it to particular tasks or domains, e.g., LLM in cybersecurity [11], Cybersecurity claim classification [12]. Additionally, researchers can explore techniques to identify and mitigate overfitting and enhance generalization and model stability by adjusting hyperparameters, regularization, and optimization methods. To make sure the model is resilient in real-world scenarios, robustness testing against adversarial attacks and domain shifts is also essential. Additionally, efforts involve concentrating on developing systems for ongoing model validation and monitoring in order to detect and address emerging biases or issues during the modeling process.
- *Post-modeling phase:* The research scope expands to deployment and governance strategies in the post-modeling phase to ensure the ethical and responsible deployment of LLMs in practical applications. In addition to comprehensive verification of the model outputs, this involves establishing mechanisms for user feedback and community engagement to promote accountability and transparency in the model deployment process. Researchers also need to give priority to creating frameworks for responsible model governance, which requires establishing explicit procedures for addressing ethical concerns, ensuring compliance with regulatory standards, and promoting responsible data management practices [13, 14]. Further efforts need to be directed towards establishing mechanisms for continuous monitoring and refinement of models to effectively address emerging ethical concerns or unintended consequences that may occur in deployed models.

By prioritizing these aspects throughout all phases, researchers can significantly contribute to the development of LLMs that not only achieve high performance but also uphold principles of trustworthiness, fairness, and responsible AI modeling throughout their lifecycle. Another crucial aspect is "prompt engineering" [1, 15] in LLMs, which involves crafting input prompts or queries to guide the model in generating desired outputs. This process is fundamental for shaping the behavior and responses of LLMs across tasks such as text generation, summarization, and question answering. Research in prompt engineering focuses on devising effective strategies to elicit specific information or responses from LLMs while minimizing biases, promoting fairness, and addressing ethical considerations. This includes exploring techniques to design clear, concise, and unambiguous prompts, as well as methods to integrate domain knowledge and contextual cues to enhance the relevance and accuracy of LLM outputs. Additionally, efforts are directed toward evaluating the impact of different prompt formulations on model performance and optimizing prompts for specific tasks or applications. Through advancements in prompt engineering techniques, researchers aim to bolster the trustworthiness, fairness, and responsible behavior of LLMs across diverse applications and domains.

## 7 Conclusion

This position paper concludes by highlighting the enormous potential of large language models while emphasizing the crucial importance of raising awareness and responsible practices in their deployment. Trustworthiness, fairness, and ethical considerations need to be given the utmost importance in the development and application of LLMs, as they continue to influence our digital world. Through advancing the principles of accountability, transparency, and interdisciplinary collaboration, stakeholders can effectively and strategically navigate the intricacies of AI modeling. We can capitalize on the transformative power of LLMs to improve human knowledge, communication, and societal well-being while mitigating potential risks and preventing unintended consequences through continued discussion, research, and innovation.

**Author contributions** The single author Dr. Iqbal H. Sarker prepared this manuscirpt.

**Data availability** No datasets were generated or analysed during the current study.

## Declarations

**Competing interests** The authors declare no competing interests.

## References

1. Zhao WX, Zhou K, Li J, et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 2023.
2. Huang L et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232, 2023.
3. Sarker IH. AI-driven cybersecurity and threat intelligence: cyber automation, intelligent decision-making and explainability. Springer, 2024.
4. Sarker IH. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. SN Comput Sci. 2021; 2(5):377.
5. Li Y et al. Large language models in finance: a survey. In Proceedings ACM Int. Conf. on AI in Finance, 2023.
6. He K et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. arXiv preprint arXiv:2310.05694, 2023.
7. Gallegos IO et al. Bias and fairness in large language models: a survey. arXiv preprint arXiv:2309.00770, 2023.
8. Waleed K, Noorhan A. Fighting lies with intelligence: using large language models and chain of thoughts technique to combat fake news. In Innovative Techniques and Applications of AI: Springer; 2023.
9. Mladjan J, Mark C. Connecting ai: Merging large language models and knowledge graph. Computer. 2023;56(11):103–8.
10. Han J et al. Data mining concepts and techniques third edition. 2012.
11. Motlagh F et al. Large language models in cybersecurity: state-of-the-art. arXiv preprint arXiv:2402.00891, 2024.
12. Kimia A, Michael H, Hamid S, Juan L, Kalyan P. Cybert: cybersecurity claim classification by fine-tuning the bert language model. J Cybersecur Privacy. 2021;1(4):615–37.
13. Dignum V. Responsible artificial intelligence: how to develop and use AI in a responsible way, volume 2156. Springer, 2019.
14. Trocin C, Mikalef P, Papamitsiou Z, Conboy K. Responsible ai for digital health: a synthesis and a research agenda. Inf Syst Front. 2023;25(6):2139–57.
15. Yamada M. Optimizing machine translation through prompt engineering: an investigation into chatgpt's customizability. arXiv preprint arXiv:2308.01391, 2023.