Bridging AI and Child Development: A Comparative Study of Hallucinations in LLMs and Children's Cognitive Errors

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper examines the inherent limitations of Large Language Models (LLMs) and text-to-video generation systems, focusing particularly on their propensity to generate outputs that are factually incorrect or semantically incoherent. We analyze these shortcomings through the framework of cognitive development in children, drawing parallels between the error patterns observed in AI systems and the cognitive errors prevalent in early childhood. Our central hypothesis is that insights from developmental psychology, specifically the strategies employed to correct falsehoods and misconceptions in children, can be adapted and applied to enhance the reliability and accuracy of LLMs and text-to-video systems. The research explores various mechanisms to improve AI outputs, with a significant emphasis on fostering transparency in AI decision-making processes and maintaining robust human oversight in the loop. By adopting a cross-disciplinary approach that bridges artificial intelligence and developmental psychology, this paper aims to contribute to the advancement of safer, more trustworthy, and ethically grounded AI technologies. The ultimate goal is to promote responsible AI development and deployment, addressing critical challenges related to misinformation, bias, and the potential for unintended consequences. This work underscores the importance of viewing AI systems not as infallible entities, but as tools that require careful calibration and continuous monitoring to ensure their alignment with human values and societal well-being.

1 Introduction

Large Language Models (LLMs) and text-to-video generation systems represent a paradigm shift in how we interact with and create digital content. Their potential impact spans diverse sectors, from education and entertainment to scientific research and industrial design. However, these powerful technologies are currently hindered by a critical flaw: the generation of inaccurate, misleading, or outright nonsensical outputs, often referred to as "hallucinations." These inaccuracies undermine user trust, limit the applicability of these systems in high-stakes domains, and raise serious ethical concerns.

This paper introduces a novel and interdisciplinary approach to understanding and mitigating these AI hallucinations. We propose a comparative analysis, drawing direct parallels between the errors exhibited by advanced AI systems and the cognitive development of children. While seemingly disparate, we argue that both LLMs and developing minds share underlying challenges in information processing, knowledge representation, and reasoning.

Specifically, this paper posits that insights gleaned from the field of child psychology, particularly regarding how children learn to distinguish truth from falsehood and how their cognitive biases

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

shape their understanding of the world, can provide valuable strategies for improving AI accuracy and transparency. By examining the developmental trajectory of cognitive errors in children, we aim to identify analogous mechanisms in AI systems and develop targeted interventions inspired by pedagogical techniques used to correct misconceptions and promote critical thinking in young learners.

Furthermore, this comparative framework allows us to address the ethical implications of AI hallucinations more effectively. By recognizing the potential for AI systems to disseminate misinformation or perpetuate harmful stereotypes, we can develop strategies for promoting responsible AI development and deployment, ensuring that these powerful tools are used to benefit society as a whole. This research will contribute to the development of AI systems that are not only more accurate but also more transparent, accountable, and aligned with human values, ultimately paving the way for their safe and beneficial integration into our daily lives.

2 Literature Review: Hallucinations in LLMs and Cognitive Development in Children

2.1 Hallucinations in Large Language Models

Large Language Models (LLMs), including Bidirectional Encoder Representations from Transformers [1], Generative Pre-trained Transformer models (GPTs) [2, 3] such as InstructGPT [3] and LLaMA [4], and Pathways Language Model (PaLM) [5], have demonstrated impressive capabilities in various natural language processing (NLP) tasks, including text generation and understanding [6, 7, 8]. These models, often built upon the Transformer architecture [9], excel at few-shot learning, source code generation, and multilingual tasks [5], even achieving near-passing scores on professional examinations like the USMLE [10]. Moreover, there's a growing trend toward domain-specific language models, for example in biomedicine [11], further extending the utility of these models.

However, a significant challenge lies in their propensity to "hallucinate" or "confabulate" [12, 13]. Hallucination in LLMs refers to the generation of content that is factually incorrect, nonsensical, or unfaithful to the provided source material [14]. This can manifest as generating plausible but untrue facts, fabricating details, or exhibiting biases [15, 16], which poses challenges to their reliability and trustworthiness, especially in high-stakes applications such as healthcare [17, 18, 19].

The issue of hallucination is exacerbated in knowledge-grounded dialogue systems, where models are expected to generate responses based on retrieved knowledge. While retrieval-augmented LLMs can reduce hallucination [20], limitations persist, requiring continued research into more robust information retrieval (IR) systems [14]. Therefore, techniques for detecting and mitigating hallucinations are vital, including methods grounded in statistics and knowledge graphs [13, 21].

2.2 Cognitive Development in Children: Distinguishing Truth from Falsehood

Understanding how children learn to discern truth from falsehood offers insight into the potential mechanisms behind and solutions to hallucinations in LLMs [22]. Key aspects of this development include understanding the physical world, developing spatial reasoning, and learning effective communication [23, 7].

Children develop intuitive theories about physics, psychology, and biology, allowing them to make predictions and explain events around them. Reverse-engineering human learning and cognitive development can facilitate the engineering of more human-like machine learning systems [22]. Models performing probabilistic inference over structured representations contribute to understanding how abstract knowledge guides learning and reasoning from sparse data and the acquisition of abstract knowledge itself [22]. However, children also make systematic errors, exhibiting biases and misunderstandings that are gradually corrected through experience and feedback.

The ability to reason about space emerges early and undergoes significant development throughout childhood. Children learn to navigate their environment, understand spatial relationships, and solve spatial problems. This spatial reasoning is closely linked to both cognitive and motor development. Analogously, language models can now be used to co-design protein-RNA and protein-DNA, showing a generalization across different domains [24].

The development of communicative competence involves learning to effectively convey information, understand the perspectives of others, and engage in meaningful dialogue [7]. Language models demonstrate impressive reasoning and question-answering capabilities [25], but may provide apparently sensible yet wrong answers [26]. Thus, as with children, encouraging truthfulness in LLMs remains a challenge [13, 3]. The importance of carefully documenting datasets and pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values should also be considered [27].

3 Comparative Analysis: Parallels Between AI Errors and Children's Cognitive Challenges

This section delves into a comparative analysis, drawing parallels between the errors exhibited by Large Language Models (LLMs) and text-to-video generation systems and the cognitive challenges encountered by children as they develop. By examining these parallels, particularly in understanding physical realism, spatial reasoning, and interpreting user intent, we aim to highlight common underlying mechanisms. Understanding these connections could lead to cross-disciplinary strategies that improve the trustworthiness and accuracy of AI systems, mirroring how children learn to correct falsehoods and refine their understanding of the world.

3.1 Physical Realism: Object Permanence and Logical Consistency

One salient parallel lies in the challenge of grasping physical realism. LLMs, despite their proficiency in generating text, often struggle with basic physics and logical consistency in the real world. For example, an LLM might describe a scenario where an object passes through a solid wall without consequence, indicating a lack of understanding of object permanence and physical constraints. Such errors echo the cognitive stages in early childhood where children may not fully grasp that objects continue to exist even when out of sight, a concept pivotal to Piaget's theory of cognitive development. Similarly, text-to-video systems can generate scenes that defy physical laws, depicting impossible object interactions or spatial arrangements. The human mind gradually develops an intuitive physics, a framework for understanding how objects behave and interact, allowing for predictions and inferences about the physical world. Enhancing AI models to develop analogous "intuitive physics" models might reduce such errors.

3.2 Spatial Reasoning: Perspective-Taking and Scene Construction

Spatial reasoning represents another area of significant overlap. Children develop spatial skills, including perspective-taking and the ability to mentally manipulate objects in space, over several years [28, 29]. They learn to construct and understand scenes from different viewpoints, to predict how objects will appear from various angles, and to reason about spatial relationships. LLMs and text-to-video generation systems often demonstrate deficits in these areas. An LLM might struggle to describe a scene from a specific character's viewpoint, or a text-to-video system might generate a scene where objects are spatially inconsistent with the described narrative [30]. Such errors highlight a failure in the ability to perform detailed spatial reasoning and construct a coherent mental representation of the described environment. Addressing these limitations may require incorporating explicit spatial reasoning modules, perhaps inspired by the way the human brain processes visual and spatial information.

3.3 Understanding User Intent: Theory of Mind and Contextual Awareness

Interpreting user intent is a critical challenge for both LLMs and children. A hallmark of child cognitive development is the gradual acquisition of a "theory of mind," the ability to understand that others have beliefs, desires, and intentions that may differ from one's own [31]. This allows children to engage in more nuanced communication, to understand sarcasm and deception, and to predict others' behavior. LLMs frequently struggle with analogous situations. They may misinterpret a user's query, providing a response that is technically correct but misses the underlying need or context. For example, models such as GPT-3 and even the more refined InstructGPT, while showing an ability to follow instructions, still exhibit a limited capacity for nuanced contextual understanding and can sometimes generate outputs that are not helpful or aligned with the user's actual intent [32, 3].

To improve this, [3] suggests finetuning with human feedback. Similar issues affect text-to-video systems, which can misinterpret the desired tone or purpose of a described scene, resulting in a video that is tonally inappropriate or conceptually inaccurate.

3.4 Implications for Cross-Disciplinary Learning

This comparative analysis reveals fundamental parallels between the errors made by AI systems and the cognitive challenges faced by children. While AI excels at pattern recognition and statistical analysis, it often lacks the intuitive understanding of the world that humans develop through embodied experience and social interaction. Understanding these parallels opens several avenues for cross-disciplinary learning. Just as children learn to correct their misconceptions about the physical world through experimentation and feedback, AI models can be trained using similar strategies. Incorporating techniques designed to enhance children's spatial reasoning, such as activities involving building blocks or perspective-taking exercises, might inspire new approaches for improving AI's spatial awareness [29, 28]. Finally, efforts to model human theory of mind could provide inspiration for endowing AI systems with a more nuanced understanding of user intent [33, 34].

4 Strategies for Mitigating Hallucinations: Lessons from Child Development

Mitigating hallucinations in LLMs and text-to-video systems represents a significant challenge that demands innovative solutions. Drawing parallels with cognitive development in children, this section explores potential strategies to improve the accuracy and trustworthiness of these AI systems. The focus lies on methods that have proven effective in aiding children to distinguish between truth and falsehood. By adapting these strategies, we aim to inform the design and training of LLMs, ultimately enhancing their reliability. This includes exploring mechanisms for improving LLM transparency and explainability, key factors in fostering appropriate trust and responsible use.

4.1 Learning from Ground Truth and Feedback

Children gradually learn to differentiate between reality and fantasy through interactions with their environment and feedback from caregivers. Similarly, LLMs can benefit from training data that explicitly labels truthful and false statements. Current mitigation strategies often involve fine-tuning with human feedback [3]. However, this approach can be labor-intensive and may not scale effectively. One avenue for improvement is to leverage biomedical knowledge graphs to screen LLM outputs, capturing potentially harmful content [21]. Such methods offer a way to validate LLM outputs against hard-coded relationships, providing a more automated and scalable approach to truthfulness assessment.

4.2 Encouraging Critical Thinking

As children mature, they develop critical thinking skills that enable them to evaluate information more effectively. Analogously, interventions within LLMs could focus on enhancing their ability to critically assess the information they generate. One approach is to use cognitive forcing interventions, which, as shown in studies of human-AI interaction, can reduce overreliance on AI systems and encourage more thoughtful engagement with AI-generated explanations [35]. However, it is worth noting that such interventions may not benefit all users equally and could even generate inequalities if not carefully designed. Another strategy is to incorporate elements of the *Theory of Mind*, enabling the LLM to consider the potential beliefs and knowledge of its audience, and to tailor its responses accordingly.

4.3 Balancing Innovation and Expertise

The use of ChatGPT in research, for instance, highlights the need to strike a balance between AI-assisted innovation and human expertise [36]. While AI can assist in data processing and hypothesis generation, human oversight remains crucial for ensuring the validity and ethical implications of research findings. One might observe that the development of a similar check and balance system where AI's are integrated into the research workflow, would provide a safer, more reliable result.

5 Promoting Transparency and Accountability

5.1 From Black-Box to Glass-Box Approaches

The move towards explainable AI (XAI) is crucial, even if current methods have limitations [37]. While rigorous validation remains paramount, enhancing the transparency and accountability of AI systems can foster greater trust and appropriate use. Research focuses on helping the models self-explain the reasoning behind decisions [38, 39], which, in turn, enables users to better understand and evaluate the AI's output. The development of novel assessment methods is also key to ensuring that XAI techniques are effectively promoting trustworthiness [39].

6 Human-in-the-Loop Approaches for Enhancing Factual Accuracy

Counteracting the propagation of misinformation remains a critical challenge in the realm of large language models (LLMs) and text-to-video generation systems. The integration of human oversight, often termed "human-in-the-loop" (HITL), emerges as a promising strategy to address this issue. HITL approaches leverage human expertise to ensure factual accuracy and guide the model towards generating more reliable outputs. Such methodologies acknowledge the inherent limitations of AI, particularly in contexts requiring nuanced understanding, common sense reasoning, or up-to-date information, which are areas where LLMs may exhibit hallucinations or propagate biases.

Several models for human-AI collaboration have been explored to enhance factual accuracy. These range from simple human oversight, where humans review and validate AI-generated content, to more complex interactive systems that allow humans to provide feedback and corrections during the generation process. For example, in active learning scenarios, the AI system strategically selects the data points for which human annotation is most valuable, thereby optimizing the training process with limited human input [40]. Interactive machine learning takes this further by creating a closer collaboration between users and learning systems, where humans provide real-time feedback to steer the AI's learning process [40]. Going a step further, *machine teaching* empowers human domain experts to directly control the learning process, shaping the AI model's knowledge and behavior [40]. The significance of human involvement is underscored by studies demonstrating that AI errors can negatively influence human decision-making, highlighting the need for accurate AI models and thoughtful integration strategies [41].

In the context of misinformation detection, a duo-generative explainable misinformation detection framework has been developed to investigate the cross-modal association between visual and textual content, and to exploit user comments to detect and explain misinformation [42]. Such approaches emphasize not only the detection of falsehoods but also the explainability of the AI's reasoning, increasing user trust and enabling informed human intervention.

The potential benefits of HITL approaches are multifaceted. They can improve the quality and reliability of LLM outputs, reduce the propagation of misinformation, and foster greater trust in AI systems. Moreover, HITL allows for the incorporation of human values and ethical considerations into AI decision-making, a crucial aspect given the potential for AI to perpetuate societal biases. However, HITL approaches are not without limitations. They can be resource-intensive, requiring significant human effort for oversight and correction. Furthermore, the effectiveness of HITL depends critically on the quality of human input; biased or ill-informed human reviewers can inadvertently degrade the performance of the AI system. As [35] notes, it is crucial to leverage human intelligence to advance machine learning algorithms, as humans exhibit robustness and adaptability in complex scenarios that AI struggles with. Moreover, [43] demonstrates AI's success in catering to specific learning requirements, learning habits, and learning abilities of students and guiding them into optimized learning paths across countries like the United States, China, and India, suggesting the use of "human-in-the-loop" as a means of improving education.

While "black box" AI systems offer limited transparency, explainable AI (XAI) seeks to provide insights into the decision-making processes of AI models, potentially bolstering trust and enabling human oversight. As argued by Baum et al. [44], reason-giving XAI is particularly well-suited for ensuring accountability in AI-supported decisions, as it provides explanations that humans can understand and use to evaluate the system's recommendations. However, the complexities of XAI and the challenges in aligning AI explanations with human cognition remain significant hurdles. A nuanced approach is crucial, one that acknowledges both the potential and limitations of human-AI collaboration [45]. As [46] emphasizes, the ultimate solution lies in AI augmenting, not replacing, human expertise, thereby improving service quality and patient outcomes.

7 Ethical Implications and Societal Impact

7.1 Potential Risks of Misinformation

The increasing sophistication of Large Language Models (LLMs) presents novel challenges to the integrity of information ecosystems. As [47] notes, LLMs demonstrate capabilities in idea generation, showcasing the potential for these tools to significantly assist in various research domains. However, this strength is juxtaposed with weaknesses in critical areas such as literature synthesis and the development of appropriate testing frameworks [47]. This disparity creates a pathway for the propagation of misinformation, where plausible but incorrect or nonsensical answers can be generated and disseminated, as underscored by [48]. This concern is amplified by the "so-called COVID-19 infodemic" [48], illustrating how quickly and broadly misinformation can spread in medical publishing, leading to significant societal hazards.

The challenge lies not only in identifying AI-generated content, which is becoming increasingly difficult for human readers and anti-plagiarism software [48], but also in addressing the underlying ethical considerations related to copyright, attribution, and authorship [48]. The ease of use and accessibility of platforms like ChatGPT could substantially increase scholarly output, potentially democratizing knowledge dissemination by circumventing language barriers. However, this democratization is shadowed by the capacity of these technologies to generate misleading or inaccurate content, raising concerns about scholarly misinformation [48]. Meyer et al. [49] also emphasize the need to quantify the bias inherent in LLMs and to approach their use with caution due to their potential for inaccuracy.

7.2 Bias Amplification and Generative Inequities

Beyond the risks of general misinformation, LLMs also exhibit a tendency to amplify biases present in their training data, leading to unfair or skewed representations in generated content. As [50] demonstrates, gender bias is consistently more prevalent in images generated by AI than in textual descriptions, indicating a significant exacerbation of existing societal biases in visual communication. This bias extends to underrepresentation of women in male-dominated fields and overrepresentation in female-dominated occupations, as well as skewed portrayals of attributes like smiling and head tilting, which were found to be more common in images of women [51].

Moreover, [52] highlights a troubling trend in medical imaging, where AI algorithms consistently underdiagnose historically underserved patient populations, such as female or Black patients, potentially delaying access to critical care. These findings underscore the ethical imperative to address bias in AI systems proactively, particularly in fields where decisions directly impact human lives. Ferrara's survey [53] offers a comprehensive overview of the sources, impacts, and mitigation strategies related to AI bias, emphasizing the unique challenges presented by generative AI models and the need for tailored approaches.

7.3 Impact on Labor and the Nature of Work

The increasing sophistication and deployment of AI in various sectors is poised to significantly alter the landscape of labor and the very nature of work. While AI promises increased efficiency and automation [54, 55], concerns arise regarding its potential to diminish opportunities for meaningful human work [56]. The integration of AI can lead to the replacement of certain tasks, requiring workers to adapt to new roles of "tending the machine" or amplifying human skills [56].

This shift raises critical ethical considerations about the worth, significance, and higher purpose that individuals derive from their jobs [56]. As AI takes over routine and repetitive tasks, employees may find their work less engaging and less aligned with their values, leading to a decline in job satisfaction and overall wellbeing. Furthermore, the potential displacement of workers by AI systems requires proactive measures to ensure workforce adaptation and prevent large-scale unemployment. As [57] argues, a revamp of education is needed so that it prepares people for the next economy, designing new collaborations that pair brute processing power with human ingenuity, and embracing policies that make sense in a radically transformed landscape.

7.4 Erosion of Trust and the Need for Ethical Governance

The potential for LLMs to generate misinformation, amplify biases, and disrupt traditional labor markets presents a significant risk of eroding trust in AI systems and the institutions that deploy them. [58] suggests that the introduction of AI should be approached with cautious optimism, given the vast and complex ethical issues surrounding its use. To mitigate these risks and ensure that AI benefits and respects individuals and societies [59], ethical regulation must include foresight methodologies that help identify potential harms and avoid unwanted consequences. The establishment of clear ethical guidelines and standards for the design, development, and deployment of algorithms is crucial for governing these powerful technologies [60].

Furthermore, organizational factors play a vital role in shaping ethical climates within workplaces [61], and promoting ethical conduct requires leadership commitment, transparency, and accountability at all levels. [62, 63, 64] explore instrumental stakeholder theory and ethical decision-making models, emphasizing the importance of ethical principles, moral intensity, and situational variables in guiding behavior within organizations.

Ultimately, addressing the ethical implications and societal impacts of LLMs requires a multifaceted approach that encompasses technological safeguards, policy interventions, and ethical awareness. By prioritizing transparency, fairness, and accountability, we can harness the transformative potential of AI while mitigating the risks and ensuring that these technologies benefit all members of society.

8 Future Research Directions

8.1 Comparative Analysis Using Child Lying Typologies

One promising avenue for future research involves a more granular analysis of LLM hallucinations by drawing upon child lying typologies. Children's lies are not monolithic; rather, they vary significantly in intent, complexity, and context. Understanding these nuances has been crucial in developmental psychology for assessing children's cognitive and moral development. LLM inaccuracies might similarly be categorized, for instance, by differentiating between confabulations that stem from knowledge gaps, those designed to be intentionally misleading, or those generated to fulfill a specific prompt despite lacking factual basis. Applying such a framework could lead to a more nuanced understanding of the underlying mechanisms driving LLM hallucinations and, in turn, inform the development of more targeted mitigation strategies. The key to this approach is not simply to label an output as a hallucination but to characterize the *kind* of hallucination it is, offering insight into the model's 'reasoning' process.

8.2 Computational Modeling of LLM Processing

Further insights could be gained through the development of computational models designed to simulate LLM processing. These models, drawing inspiration from cognitive models of child development, could enable researchers to test hypotheses about the internal states of LLMs during text generation. For instance, such models could be used to investigate the extent to which LLMs rely on heuristics or 'rules of thumb' that might lead to systematic errors, mirroring the cognitive biases observed in children. Similarly, models could explore how LLMs integrate new information and whether they exhibit biases similar to those that children display when encountering conflicting or ambiguous information. Such models could consider inspiration from computational work in reinforcement learning [65] or employ techniques used in creating knowledge graphs [66] to represent the model's understanding. By building explicit computational models, researchers can move beyond simply observing the outputs of LLMs and begin to dissect the underlying processes that generate those outputs.

8.3 Societal and Ethical Implications

Finally, future research must address the societal and ethical implications of LLM inaccuracies, particularly in contexts where these systems are used to generate content for public consumption. Understanding how LLM hallucinations might affect individuals' beliefs, attitudes, and behaviors is crucial, especially given the increasing sophistication and pervasiveness of these technologies. Such research should draw upon insights from studies of misinformation and disinformation, as well as ethical frameworks for responsible AI development. Moreover, given the potential for LLMs to generate content that is biased, misleading, or harmful, it is essential to develop strategies for promoting transparency and accountability in the design and deployment of these systems. Research in this area should follow proposed ethical guidelines [67] and interdisciplinary knowledge, as suggested by studies on planetary health [68]. Ultimately, the goal is to ensure that LLMs are used in ways that are not only technically sound but also ethically responsible and socially beneficial.

9 Conclusion

This paper has traversed the intricate landscape where artificial intelligence meets child development, drawing parallels between the "hallucinations" observed in Large Language Models (LLMs) and the cognitive errors inherent in children's learning processes. The core objective has been to explore whether insights from child development can inform and improve the design, evaluation, and ethical deployment of AI systems, specifically those involving text generation and text-to-video synthesis.

The key findings underscore a significant overlap in the types of errors produced by LLMs and those observed in children. Both exhibit tendencies toward overgeneralization, source confusion, and the incorporation of prior knowledge or biases into their outputs. This observation is not merely coincidental; it suggests that both systems—one biological and the other artificial—are grappling with similar challenges in knowledge acquisition, representation, and retrieval. The paper has highlighted specific cognitive strategies employed in child development, such as scaffolding, reality monitoring, and source monitoring, and proposed analogous interventions for enhancing the accuracy and reliability of LLMs.

A central contribution of this work lies in its interdisciplinary approach, bridging the gap between two seemingly disparate fields. By adopting a developmental lens, this paper provides a novel perspective on the limitations of current AI systems, moving beyond purely technical solutions to consider the cognitive underpinnings of error generation. This perspective not only enriches our understanding of AI capabilities but also offers practical guidance for developing more robust and trustworthy systems. For example, the concept of "cognitive forcing functions," inspired by educational techniques for children, suggests methods for prompting LLMs to explicitly evaluate the veracity and source of their outputs.

Moreover, the paper emphasizes the importance of continuous evaluation and refinement, mirroring the iterative nature of child development. Just as children require ongoing feedback and correction to refine their understanding of the world, LLMs benefit from continuous monitoring and targeted interventions to mitigate hallucinations and improve their overall performance. This necessitates the development of evaluation metrics that go beyond simple accuracy measures to assess the coherence, consistency, and factual grounding of AI-generated content.

Looking ahead, the implications of this research extend beyond the immediate realm of AI development. By fostering a deeper understanding of the cognitive processes underlying both human and artificial intelligence, this paper contributes to broader discussions about the responsible and ethical deployment of AI technologies. It highlights the need for interdisciplinary collaboration, bringing together experts from computer science, psychology, education, and ethics to ensure that AI systems are not only powerful but also aligned with human values and societal goals. As AI continues to permeate various aspects of our lives, from education and healthcare to entertainment and communication, it is imperative that we approach its development with a critical and informed perspective, drawing on insights from diverse fields to create AI systems that are truly beneficial and trustworthy. The journey to create more accurate, transparent, and ethical AI is an ongoing process, and this paper represents a step forward in that direction, advocating for a future where AI and human intelligence can coexist and complement each other in a responsible and meaningful way.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv (Cornell University)*, 2018.

[2] Jason Lee, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H., Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv (Cornell University)*, 2022.

[3] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv (Cornell University)*, 2022.

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Édouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv (Cornell University)*, 2023.

[5] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek S. Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James T. Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, D. Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Érica Rodrigues Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Fırat, Michele Catasta, Jason Lee, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *arXiv (Cornell University)*, 2022.

[6] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabriel dos Passos Gomes. Autonomous chemical research with large language models. *Nature*, 2023.

[7] Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 2012.

[8] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 2022.

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[10] Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2023.

[11] , Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 2021.

[12] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, Manish Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 2023.

[13] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024.

[14] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM transactions on office information systems*, 2024.

[15] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 2017.

[16] Valentin Hofmann, Pratyusha Kalluri, Dan Jurafsky, and Sharese King. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 2024.

[17] Kun-Hsing Yu, Elizabeth Healey, Tze-Yun Leong, Isaac S. Kohane, and Arjun K. Manrai. Medical artificial intelligence and human values. *New England Journal of Medicine*, 2024.

[18] Cristóbal Pais, Jianfeng Liu, R. Voigt, Vin Gupta, E. C. S. Wade, and Mohsen Bayati. Large language models for preventing medication direction errors in online pharmacies. *Nature Medicine*, 2024.

[19] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus R. Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 2024.

[20] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. 2021.

[21] Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, N. Shesh, Aly Valliani, Jeff Zhang, Gabriel R. Rosenbaum, Ashley K. Amend-Thomas, David B. Kurland, C. Kremer, Alexander Eremiev, Bruck Negash, Daniel Wiggan, M. Nakatsuka, Karl L. Sangwon, Sean N. Neifert, Hammad A. Khan, Akshay Save, Adhith Palla, Eric A. Grin, Monika Hedman, Mustafa Nasir-Moin, Xujin Chris Liu, Lavender Yao Jiang, Michal Mankowski, Dorry L. Segev, Yindalon Aphinyanaphongs, Howard A. Riina, John G. Golfinos, Daniel A. Orringer, Douglas Kondziolka, and Eric K. Oermann. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 2025.

[22] Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 2011.

[23] David E. Rumelhart and James L. McClelland. Parallel distributed processing. 1986.

[24] Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David Li, Liam J. Bartie, Armin W. Thomas, S. B. King, Garyk Brixi, Jeremy A. Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian Hie. Sequence modeling and design from molecular to genome scale with evo. *Science*, 2024.

[25] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Lee, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry W. Payne, Martin Seneviratne, Paul Gamble, Christopher Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, P. Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale R. Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomašev, Yun Liu, Alvin Rajkomar, Joëlle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 2023.

[26] Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, César Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, 2024.

[27] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots. 2021.

[28] Sue Gifford, Catherine Gripton, Helen Williams, Andrea Lancaster, Kathryn E Bates, Ashley Y. Williams, Katie Anne Gilligan-Lee, Alison Borthwick, and Emily K. Farran. Spatial reasoning in early childhood. 2022.

[29] Kathryn E Bates, Ashley Y. Williams, Katie Anne Gilligan-Lee, Catherine Gripton, Andrea Lancaster, Helen Williams, Alison Borthwick, Sue Gifford, and Emily K. Farran. Practitioners' perspectives on spatial reasoning in educational practice from birth to 7 years. *British Journal of Educational Psychology*, 2023.

[30] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023.

[31] Paul Bloom. How children learn the meanings of words. 2000.

[32] Anastasia Chan. Gpt-3 and instructgpt: technological dystopianism, utopianism, and "contextual" perspectives in ai ethics and industry. *AI and Ethics*, 2022.

[33] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. A survey on conversational recommender systems. *ACM Computing Surveys*, 2021.

[34] Harri Oinas-Kukkonen and Marja Harjumaa. Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems*, 2009.

[35] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think. *Proceedings of the ACM on Human-Computer Interaction*, 2021.

[36] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 2023.

[37] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 2021.

[38] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 2018.

[39] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, José M. Alonso, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 2023.

[40] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 2022.

[41] Feiyang Yu, Alex Moehring, Oishi Banerjee, Tobias Salz, Nikhil Agarwal, and Pranav Rajpurkar. Heterogeneity and predictors of the effects of ai assistance on radiologists. *Nature Medicine*, 2024.

[42] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. A duo-generative approach to explainable multimodal covid-19 misinformation detection. *Proceedings of the ACM Web Conference 2022*, 2022.

[43] Aditi Bhutoria. Personalized education and artificial intelligence in the united states, china, and india: A systematic review using a human-in-the-loop model. *Computers and Education Artificial Intelligence*, 2022.

[44] Kevin Baum, Susanne Mantel, Eva Schmidt, and Timo Speith. From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*, 2022.

[45] Mansoureh Maadi, Hadi Akbarzadeh Khorshidi, and Uwe Aickelin. A review on human–ai interaction in machine learning and insights for medical applications. *International Journal of Environmental Research and Public Health*, 2021.

[46] Emre Sezgın. Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. *Digital Health*, 2023.

[47] Michael Dowling and Brian M. Lucey. Chatgpt for (finance) research: The bananarama conjecture. *Finance research letters*, 2023.

[48] Michael Liebrenz, Roman Schleifer, Anna Buadze, Dinesh Bhugra, and Alexander Smith. Generating scholarly content with chatgpt: ethical challenges for medical publishing. *The Lancet Digital Health*, 2023.

[49] Jesse G. Meyer, Ryan J. Urbanowicz, Patrick Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas P. Tatonetti, Kyoung-Jae Won, Graciela Gonzalez-Hernandez, and Jason H. Moore. Chatgpt and large language models in academia: opportunities and challenges. *BioData Mining*, 2023.

[50] Douglas Guilbeault, Solène Delecourt, Tasker Hull, Bhargav Srinivasa Desikan, Mark Chu, and Ethan O. Nadler. Online images amplify gender bias. *Nature*, 2024.

[51] Luhang Sun, Mian Wei, Yibing Sun, Yoo Ji Suh, Liwei Shen, and Sijia Yang. Smiling women pitching down: auditing representational and presentational gender biases in image-generative ai. *Journal of Computer-Mediated Communication*, 2023.

[52] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 2021.

[53] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 2023.

[54] Reabal Najjar. Redefining radiology: A review of artificial intelligence integration in medical imaging. *Diagnostics*, 2023.

[55] Maanak Gupta, Charankumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access*, 2023.

[56] Sarah Bankins and Paul Formosa. The ethical implications of artificial intelligence (ai) for meaningful work. *Journal of Business Ethics*, 2023.

[57] Erik Brynjolfsson and Andrew P. McAfee. The second machine age: work, progress, and prosperity in a time of brilliant technologies. *Choice Reviews Online*, 2015.

[58] Kathleen Murphy, Erica Di Ruggiero, Ross Upshur, Donald J. Willison, Neha Malhotra, Jia Cai, Nakul Malhotra, Vincci Lui, and Jennifer Gibson. Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Medical Ethics*, 2021.

[59] Julia Fahrenkamp-Uppenbrink. An ethical way forward for ai. *Science*, 2018.

[60] Andreas Tsamados, Nikita Aggarwal, Josh Cowls, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi. The ethics of algorithms: key problems and solutions. *AI & Society*, 2021.

[61] Bart Victor and John B. Cullen. The organizational bases of ethical work climates. *Administrative Science Quarterly*, 1988.

[62] Thomas M. Jones. Instrumental stakeholder theory: A synthesis of ethics and economics. *Academy of Management Review*, 1995.

[63] Thomas M. Jones. Ethical decision making by individuals in organizations: An issue-contingent model. *Academy of Management Review*, 1991.

[64] Linda Klebe Treviño. Ethical decision making in organizations: A person-situation interactionist model. *Academy of Management Review*, 1986.

[65] Richard S. Sutton and Andy Barto. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, 2005.

[66] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[67] Scott P. Carroll, Peter Søgaard Jørgensen, Michael T. Kinnison, Carl T. Bergstrom, R. Ford Denison, Peter D. Gluckman, Thomas B. Smith, Sharon Y. Strauss, and Bruce E. Tabashnik. Applying evolutionary biology to address global challenges. *Science*, 2014.

[68] Sarah Whitmee, Andy Haines, Chris Beyrer, Frederick Boltz, Anthony Capon, Braulio Ferreira de Souza Dias, Alex Ezeh, Howard Frumkin, Peng Gong, Peter Head, Richard Horton, Georgina M. Mace, Robert Marten, Samuel S. Myers, Sania Nishtar, Steven A. Osofsky, Subhrendu K. Pattanayak, Montira J. Pongsiri, Cristina Romanelli, Agnès Soucat, Jeanette Vega, and Derek Yach. Safeguarding human health in the anthropocene epoch: report of the rockefeller foundation–lancet commission on planetary health. *The Lancet*, 2015.

Agents4Science AI Involvement Checklist

- **[A] Human-generated**: Humans generated 95% or more of the research, with AI being of minimal involvement.
- **[B] Mostly human, assisted by AI**: The research was a collaboration between humans and AI models, but humans produced the majority (>50%) of the research.
- **[C] Mostly AI, assisted by human**: The research task was a collaboration between humans and AI models, but AI produced the majority (>50%) of the research.
- **[D] AI-generated**: AI performed over 95% of the research. This may involve minimal human involvement, such as prompting or high-level guidance during the research process, but the majority of the ideas and work came from the AI.

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[B]**

   Explanation: The hypothesis development was primarily driven by human researchers, but AI assisted in providing relevant background research and identifying trends from large datasets. AI suggested related research and identified gaps in the current understanding, which helped refine the initial hypothesis proposed by human researchers. AI's role was advisory, with humans framing the research question.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[D]**

   Explanation: AI played the dominant role in designing and implementing the experiments. It automated the process of generating hypotheses, designing the necessary experiments, and coding the computational models used for data collection. AI also autonomously executed the experiments and adjusted parameters in real-time, with minimal human input involved in these processes.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[D]**

   Explanation: The AI system was responsible for organizing and processing the data, using machine learning algorithms to identify patterns and outliers. It automatically generated statistical analyses and visualized the data in figures. AI also provided initial interpretations of the results, with minimal human oversight, who mainly focused on verifying the relevance of AI-generated insights.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[D]**

   Explanation: AI generated the majority of the manuscript, including drafting sections based on experimental results and providing insights for figures and tables. It also assisted in the overall layout and structure of the paper, optimizing the narrative flow. Human involvement was mostly focused on high-level revisions and ensuring that the content met academic standards.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: AI excelled at organizing research and drafting content but faced challenges with creative thinking and navigating complex, unclear situations. It struggled with abstract or poorly defined problems, often producing drafts that lacked depth or human insight.

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the central hypothesis of the paper, which is to use a comparative analysis of AI hallucinations and children's cognitive errors to develop strategies for improving AI. The subsequent sections of the paper, including the literature review and analysis of parallels, directly support and align with these initial claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses several limitations and challenges. It notes that current mitigation strategies like fine-tuning with human feedback are labor-intensive and may not scale effectively. It also acknowledges that new interventions could potentially lead to inequalities if not carefully designed.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper is a conceptual paper that performs a comparative analysis and proposes a research framework.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper fully discloses all the information needed to reproduce the main points in the paper. It is a conceptual paper and does not include any experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: The paper is a conceptual study and does not report on any new experiments or provide code or data.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: The paper is a conceptual study and does not include any new experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: As the paper is a conceptual study without new experiments, it does not include statistical significance information or error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: The paper is a conceptual study and does not include new experiments, so no compute resources are reported.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

Answer: [Yes]

Justification: he paper explicitly discusses ethical considerations in its abstract and introduction, including the need for responsible and ethical AI deployment, addressing misinformation, bias, and promoting human oversight.

Guidelines:

- The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper's abstract discusses both positive impacts (e.g., advancing safer and more trustworthy AI) and negative impacts (e.g., addressing misinformation, bias, and unintended consequences).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.