A Survey of Hybrid Human-Artificial Intelligence and Ethical Considerations

www.surveyx.cn

Abstract

This survey paper examines the interdisciplinary integration of Hybrid Human-Artificial Intelligence, focusing on the synergy between human intelligence and advanced AI technologies like large language models, deep learning, reinforcement learning, and natural language processing. The paper is structured to provide a comprehensive overview, beginning with foundational concepts and theoretical frameworks that support the development of hybrid systems. It explores the transformative potential of these systems across therapeutic and creative domains, highlighting their ability to enhance decision-making and foster innovation. Key sections delve into the integration of knowledge graphs within large language models, optimizing performance and cost management, and the challenges and advancements in reinforcement learning for natural language processing tasks. The dynamics of human-AI collaboration are analyzed, emphasizing frameworks and models that facilitate effective interaction and the importance of trust and transparency. Ethical considerations are addressed, focusing on bias, fairness, accountability, and the development of ethical guidelines and standards. The conclusion underscores the transformative potential of hybrid human-AI systems and the need for ongoing ethical considerations, offering future directions for research and development. This survey highlights the importance of balancing technological advancements with ethical imperatives to ensure responsible and beneficial AI deployment.

1 Introduction

1.1 Structure of the Survey

This survey provides a comprehensive examination of Hybrid Human-Artificial Intelligence (H-AI) and its ethical implications, focusing on the integration of human intelligence with AI to tackle complex social challenges and the ethical considerations arising from their interaction in domains such as scholarly peer review and social computing [1, 2, 3, 4, 5]. The **Introduction** establishes the context by emphasizing the synergy between human intelligence and advanced AI technologies, including large language models, deep learning, reinforcement learning, and natural language processing, and underscores the significance of human-AI collaboration along with associated ethical concerns.

The **Background and Definitions** section presents a thorough overview of essential concepts, defining key terms such as Hybrid Human-Artificial Intelligence, Large Language Models, Deep Learning, Reinforcement Learning, Natural Language Processing, Human-AI Collaboration, and Ethical AI, while also discussing the evolution and current state of these technologies.

In the section titled **Theoretical Foundations and Frameworks**, the survey explores the theoretical principles and conceptual frameworks guiding the design and implementation of H-AI systems, illustrating how these systems merge human cognitive abilities with AI to address complex social issues more effectively than traditional AI approaches, thereby laying a robust foundation for advancing social computing technologies [4, 6, 2, 3].

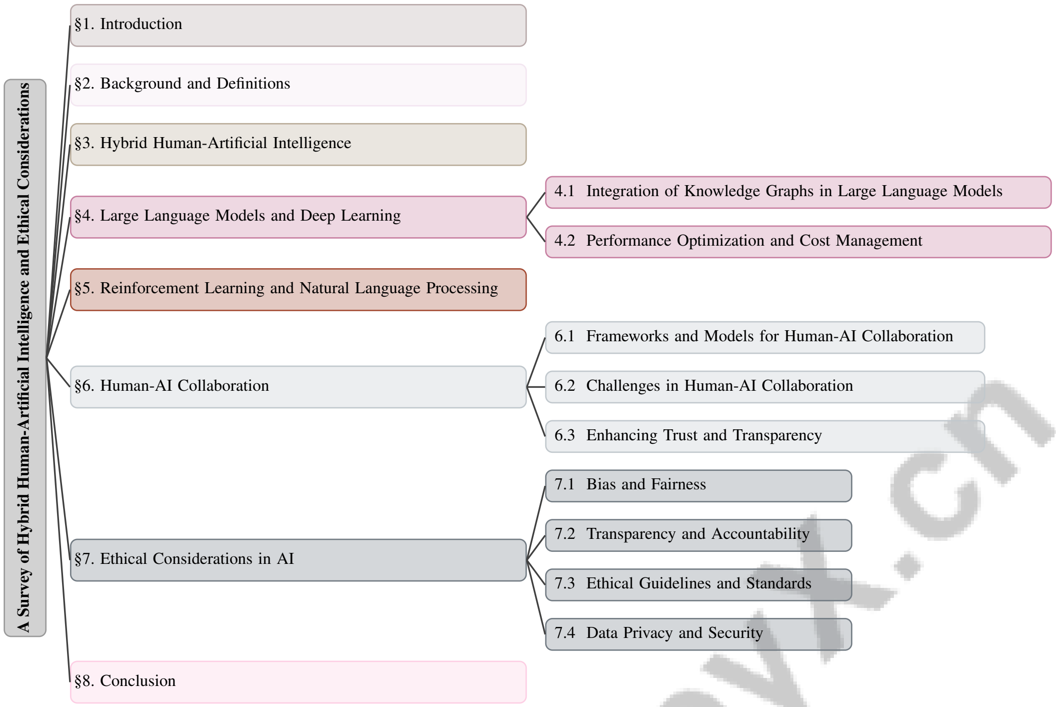

A Survey of Hybrid Human-Artificial Intelligence and Ethical Considerations
§1. Introduction
§2. Background and Definitions
§3. Hybrid Human-Artificial Intelligence
§4. Large Language Models and Deep Learning
4.1 Integration of Knowledge Graphs in Large Language Models
4.2 Performance Optimization and Cost Management
§5. Reinforcement Learning and Natural Language Processing
§6. Human-AI Collaboration
6.1 Frameworks and Models for Human-AI Collaboration
6.2 Challenges in Human-AI Collaboration
6.3 Enhancing Trust and Transparency
§7. Ethical Considerations in AI
7.1 Bias and Fairness
7.2 Transparency and Accountability
7.3 Ethical Guidelines and Standards
7.4 Data Privacy and Security
§8. Conclusion

Figure 1: chapter structure

The section on **Hybrid Human-Artificial Intelligence** investigates the concept of hybrid intelligence, where human cognitive strengths are combined with AI systems to enhance decision-making and problem-solving capabilities, supported by various examples and applications across different fields.

The subsequent section, **Large Language Models and Deep Learning**, assesses the influence of these technologies on advancing AI capabilities, particularly their role in natural language processing and integration into hybrid systems.

The study titled **"Reinforcement Learning and Natural Language Processing"** provides a detailed analysis of the application of reinforcement learning (RL) techniques in various natural language processing (NLP) tasks, especially within conversational systems. It highlights the potential of RL algorithms, particularly those employing deep neural networks, to enhance AI systems' sophistication in understanding and generating human language. Additionally, the paper addresses challenges related to reward model specification in RL, including alignment with user values and limitations of feedback from large language models. This exploration contributes to the ongoing development of AI systems that more effectively integrate human intelligence and reasoning capabilities, thereby facilitating improved NLP applications [7, 8, 9, 10].

The dynamics of **Human-AI Collaboration** are examined through various frameworks and models that enhance interaction, emphasizing the importance of transparency and explainability in AI systems. This investigation highlights potential benefits, such as improved learning outcomes and increased user trust, alongside challenges related to ethical considerations and the complexities of AI integration across diverse contexts. By prioritizing human-centered approaches, the research identifies actionable strategies to promote responsible AI development and implementation in areas like education and peer review processes [11, 12, 13, 14, 5].

The section titled **Ethical Considerations in AI** explores the multifaceted ethical implications of integrating AI technologies within hybrid systems, focusing on critical issues such as bias, transparency, and accountability. It further discusses the unique ethical challenges posed by Large Language Models (LLMs), including concerns about privacy, fairness, hallucination, and the complexities of verifiable accountability and censorship. The discussion advocates for the establishment of tailored ethical guidelines and dynamic auditing systems to address these challenges, promoting responsible AI deployment while ensuring alignment with moral and epistemic norms across various contexts, particularly in sensitive domains such as scholarly peer review [5, 15].

AI-generated, for reference only.

The **Conclusion** synthesizes key findings, highlighting the transformative potential of hybrid human-AI systems and the critical importance of ethical considerations in their development and deployment. The section on **Future Directions and Ethical Considerations** outlines prospective research trajectories and ongoing ethical challenges in this evolving field.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Theoretical Foundations and Frameworks

Hybrid human-AI systems are underpinned by theoretical frameworks that integrate cognitive science with AI, enhancing decision-making processes. Large language models (LLMs) and multimodal large language models (MLLMs) are classified into encoder-only, decoder-only, and encoder-decoder architectures, each offering unique capabilities [16]. This classification is crucial for optimizing decision-making and fostering human-AI synergies.

Hybrid systems often merge symbolic and sub-symbolic techniques, employing sequential, nested, cooperative, and compiled architectures to integrate symbolic reasoning with sub-symbolic learning [8]. Path-based relational reasoning over knowledge graphs is significant for question answering and recommender systems, facilitating complex information processing akin to human cognition [17].

The Language-Goal-Behavior (LGB) architecture decouples skill acquisition from language grounding through semantic representations, enhancing human-AI communication [18]. Reinforcement Learning from Human Feedback (RLHF), as seen in ChatGPT, demonstrates AI adaptability by refining outputs based on user feedback [19]. This adaptability, combined with supervised learning, improves applications like image caption generation, tailoring AI systems to user needs [20].

The dual intelligence perspective emphasizes collaboration between human and AI capabilities, where AI augments human decision-making rather than replacing it [2]. Frameworks categorizing deceptive patterns in AI design highlight ethical considerations, promoting transparency and accountability in human-AI collaborations [21].

Modeling human evaluators as Boltzmann-rational agents reveals how partial observations influence AI learning, providing insights into cognitive biases affecting AI evaluations [22]. Understanding these biases is vital for developing AI systems that accurately assess and adapt to human feedback.

Frameworks addressing communication challenges—such as knowledge, timing, and emotional gaps—emphasize AI's role in bridging these divides [23]. Integrating Differentiable Fuzzy Logics (DFL) with gradient-based learning optimizes neurosymbolic systems, enhancing hybrid models' adaptability and efficiency [24].

The critique of epistemic monoculture in AI research highlights the need for a balanced approach that values diverse epistemic criteria beyond predictive accuracy [25]. The interplay between Human Behavioral Ecology (HBE) and reinforcement learning (RL) offers a framework for exploring dynamic human and AI learning processes, advancing hybrid systems' capabilities [26]. Contributions from naturalistic mental representation [27] and cognitive psychology [28] further enrich these theoretical foundations, providing a comprehensive basis for sophisticated hybrid system development.

In recent years, the convergence of human and artificial intelligence has led to significant advancements in both therapeutic and creative domains. As illustrated in Figure 2, this figure depicts the integration of hybrid human-artificial intelligence systems, showcasing their synergies in complex decision-making processes. The hierarchical structure not only highlights the various applications and innovations within these fields but also emphasizes the role of AI in enhancing human experiences while addressing intricate challenges. Furthermore, it outlines the collaborative frameworks that align AI actions with human values, thereby underscoring the critical importance of ethical considerations and human oversight in AI-driven processes. This visual representation serves to reinforce the narrative of our discussion, providing a clear and engaging reference point for understanding the multifaceted interactions between human and artificial intelligence.

AI-generated, for reference only.

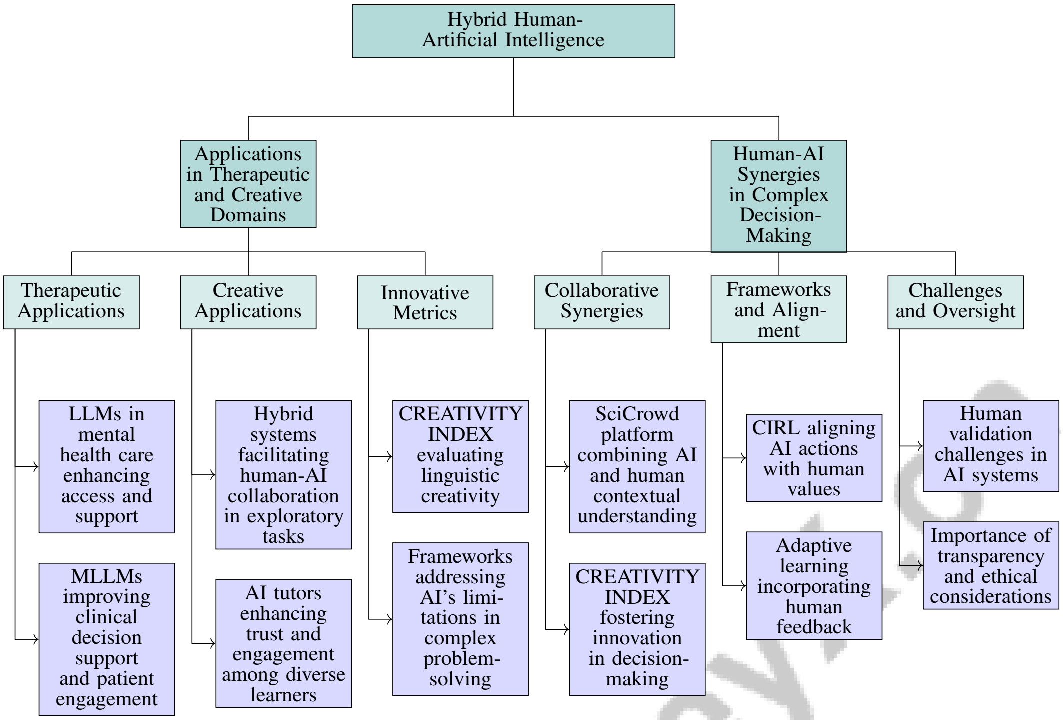

Hybrid Human-Artificial Intelligence
Applications in Therapeutic and Creative Domains
Human-AI Synergies in Complex Decision-Making
Therapeutic Applications
Creative Applications
Innovative Metrics
Collaborative Synergies
Frameworks and Alignment
Challenges and Oversight
LLMs in mental health care enhancing access and support
MLLMs improving clinical decision support and patient engagement
Hybrid systems facilitating human-AI collaboration in exploratory tasks
AI tutors enhancing trust and engagement among diverse learners
CREATIVITY INDEX evaluating linguistic creativity
Frameworks addressing AI's limitations in complex problem-solving
SciCrowd platform combining AI and human contextual understanding
CREATIVITY INDEX fostering innovation in decision-making
CIRL aligning AI actions with human values
Adaptive learning incorporating human feedback
Human validation challenges in AI systems
Importance of transparency and ethical considerations

Figure 2: This figure illustrates the integration of hybrid human-artificial intelligence systems in therapeutic and creative domains, as well as their synergies in complex decision-making. The hierarchical structure depicts the applications and innovations in therapeutic and creative fields, highlighting the role of AI in enhancing human experiences and addressing complex challenges. It also outlines the collaborative synergies and frameworks that align AI actions with human values, emphasizing the importance of ethical considerations and human oversight in AI-driven processes.

3 Hybrid Human-Artificial Intelligence

3.1 Applications in Therapeutic and Creative Domains

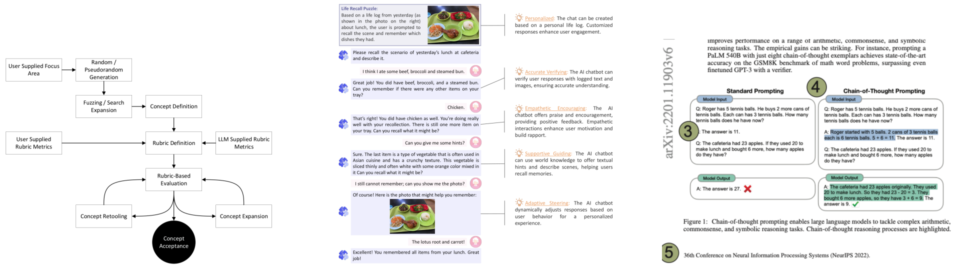
Hybrid human-AI systems are revolutionizing therapeutic and creative domains by leveraging AI technologies to enhance outcomes and foster innovation. In mental health care, large language models (LLMs) are pivotal in overcoming barriers to access, as demonstrated by the HELPERT model, which is benchmarked against human counselors in delivering Cognitive Behavioral Therapy (CBT)-based support [29]. These models are also evaluated for comprehending human expressions of mental health conditions, increasing their real-world applicability [30].

In healthcare, multimodal large language models (MLLMs) enhance clinical decision support, medical imaging, and patient engagement, leading to improved diagnostic accuracy and personalized treatment plans [31]. In creative fields, hybrid systems facilitate human-AI collaboration, as seen in the CARE system, which aids users in exploratory tasks like travel planning and skill acquisition [32]. AI tutors, incorporating cultural knowledge tailored to specific communities, enhance trust and engagement among diverse learners, including the Deaf or Hard of Hearing (DHH) [33].

The CREATIVITY INDEX offers a scalable metric for evaluating linguistic creativity, challenging conventional methods and promoting innovative applications across domains [1]. Hybrid human-AI systems in therapeutic and creative fields not only address AI's limitations in complex problem-solving but also enrich human experiences, providing innovative solutions to multifaceted challenges. These frameworks effectively tackle social issues, while advancements like the CREATIVITY INDEX reveal differences between human and machine-generated content, expanding the potential for social computing and creative expression [4, 2, 1].

As depicted in Figure 3, the integration of human creativity and machine intelligence in AI systems introduces new possibilities in therapeutic and creative domains. The "Conceptual Framework for

AI-generated, for reference only.



Life Recall Puzzle:
Based on a life log from yesterday (as shown in the photo on the right) about lunch, the user is prompted to recall the scene and remember which dishes they had.

Please recall the scenario of yesterday's lunch at cafeteria and describe it.

I think I ate some beef, broccoli and steamed bun.

Great job! You did have beef, broccoli, and a steamed bun. Can you remember if there were any other items on your tray?

Chicken.

That's right! You did have chicken as well. You're doing really well with your recollection. There is still one more item on your tray. Can you recall what it might be?

Can you give me some hints?

Sure. The last item is a type of vegetable that is often used in Asian cuisine and has a crunchy texture. This vegetable is sliced thinly and often white with some orange color mixed in it Can you recall what it might be?

I still cannot remember; can you show me the photo?

Of course! Here is the photo that might help you remember:

The lotus root and carrot!

Excellent! You remembered all items from your lunch. Great job!

Personalized: The chat can be created based on a personal life log. Customized responses enhance user engagement.

Accurate Verifying: The AI chatbot can verify user responses with logged text and images, ensuring accurate understanding.

Empathetic Encouraging: The AI chatbot offers praise and encouragement, providing positive feedback. Empathetic interactions enhance user motivation and build rapport.

Supportive Guiding: The AI chatbot can use world knowledge to offer textual hints and describe scenes, helping users recall memories.

Adaptive Steering: The AI chatbot dynamically adjusts responses based on user behavior for a personalized experience.

arXiv:2201.11903v6

improves performance on a range of arithmetic, commonsense, and symbolic reasoning tasks. The empirical gains can be striking. For instance, prompting a PaLM 540B with just eight chain-of-thought exemplars achieves state-of-the-art accuracy on the GSM8K benchmark of math word problems, surpassing even finetuned GPT-3 with a verifier.

Standard Prompting — Model Input:
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output:
A: The answer is 27. ✗

Chain-of-Thought Prompting — Model Input:
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output:
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

(a) Conceptual Framework for Evaluating and Expanding Concepts[34]

(b) Life Recall Puzzle[35]

(c) Chain-of-Thought Prompting Enhances Large Language Models for Complex Arithmetic, Commonsense, and Symbolic Reasoning Tasks[36]

Figure 3: Examples of Applications in Therapeutic and Creative Domains

Evaluating and Expanding Concepts" demonstrates a structured approach to concept development, where generative and evaluative steps refine user input, highlighting AI's collaborative role in enhancing human conceptualization processes. The "Life Recall Puzzle" uses AI to aid memory recall, offering cognitive reinforcement by prompting users to reconstruct past experiences. The "Chain-of-Thought Prompting" approach enhances large language models' capabilities in complex reasoning tasks by structuring prompts to mimic human thought processes, bridging human intuition and machine computation. These examples underscore the promising applications of hybrid human-AI systems in expanding creativity and therapy boundaries [34, 35, 36].

3.2 Human-AI Synergies in Complex Decision-Making

The integration of human and AI capabilities in complex decision-making processes is crucial for achieving enhanced outcomes through collaborative synergies. These synergies exploit the strengths of both entities, enabling nuanced and accurate solutions. The SciCrowd platform exemplifies this integration by combining AI's computational power with human contributors' contextual understanding, facilitating precise scientometric analyses [3]. This highlights AI's potential to augment human decision-making by providing deeper insights and expanding the solution space.

In creative domains, metrics like the CREATIVITY INDEX assess Large Language Models (LLMs), showcasing AI's ability to generate novel ideas while leveraging human knowledge [1]. This metric fosters innovation and enhances decision-making processes by providing a framework for evaluating and improving AI-generated solutions.

Frameworks such as Cooperative Inverse Reinforcement Learning (CIRL), combined with cognitive models of human decision-making, offer practical approaches to aligning AI actions with human values and intentions. This alignment is essential for ensuring AI systems complement human decision-making processes, enhancing outcomes in collaborative environments. By incorporating human feedback, these frameworks facilitate adaptive learning that aligns AI actions with human objectives and ethical principles, addressing biases and transparency in AI-mediated applications like peer review and collaborative robotics. Such alignment is vital for maintaining scholarly practices' integrity and enhancing trust in human-AI interactions, ensuring AI systems reflect the scholarly community's moral and epistemic norms [5, 37, 14, 38].

Challenges associated with human validation in AI systems, such as potential bias and the need for specialized domain knowledge, underscore the necessity for robust human oversight. These issues threaten the integrity and credibility of AI-driven processes like peer review, highlighting the importance of transparency and ethical considerations in deploying AI technologies. Understanding these dynamics is essential for ensuring the alignment between human expertise and AI capabilities, safeguarding outcomes' quality and reliability across various fields [39, 11, 40, 5, 41]. Addressing these challenges is crucial for enhancing AI-assisted decision-making quality, ensuring human oversight remains unbiased and effective.

The integration of human and AI capabilities in complex decision-making processes holds substantial potential to enhance outcomes across diverse fields, as evidenced by tools like SciCrowd, which facilitates collaborative scientometric analysis, and AI-Interpret, which transforms opaque decision-making policies into interpretable strategies, ultimately improving human decision-making effectiveness [1, 42, 3]. By leveraging the strengths of both humans and AI, these collaborative efforts can transform decision-making processes, leading to more innovative and effective solutions.

4 Large Language Models and Deep Learning

Large language models (LLMs) have transformed natural language processing by generating coherent and contextually relevant text. Their capabilities are significantly enhanced through the integration of knowledge frameworks, particularly knowledge graphs (KGs), which offer structured representations of information. This integration improves reasoning, interaction capabilities, and the understanding of complex data structures. The following subsection explores methods of incorporating KGs within LLMs and their effects on model performance and versatility.

4.1 Integration of Knowledge Graphs in Large Language Models

The integration of knowledge graphs (KGs) into large language models (LLMs) marks a pivotal advancement in enhancing AI systems' reasoning and interaction capabilities. KGs provide structured, semantic-rich representations that, when combined with LLMs, enhance the models' ability to process and reason with complex data. This integration can occur at various stages: before-training, during-training, and post-training, each contributing to the robust development of LLMs [43].

Before-training integration involves embedding KGs to establish a foundational knowledge base, enhancing LLMs' contextual reasoning from the outset. During-training enhancement utilizes KGs to dynamically refine the models' knowledge, enabling nuanced understanding of diverse data inputs, particularly useful in complex domains [44]. Post-training integration focuses on fine-tuning LLMs with KGs, addressing gaps in reasoning and interaction capabilities, and adapting models to specific tasks and domains, thereby improving robustness and reliability [45].

Advanced architectures, such as the MRKL system, combine language models with discrete reasoning modules to handle a wider range of tasks, illustrating the potential of KGs for improved performance [46]. Additionally, GraphLLM processes graph structures directly, enhancing reasoning capabilities without converting data to text [47]. The neuro-symbolic approach further integrates external knowledge through KGs, enriching LLMs' abstraction and mapping processes, thus improving interpretative and reasoning capabilities [12]. Incorporating KGs into multimodal large language models (MLLMs) enhances multimodal reasoning, facilitating more natural interactions with users [48].

In practical applications, KGs enable LLMs to achieve human-like communication and adaptive decision-making, particularly in autonomous systems. For instance, integrating KGs in autonomous vehicles enhances contextual understanding and reliability in dynamic environments [49]. Frameworks like AgentSense exemplify the enhancement of LLMs in modeling complex social interactions, improving their ability to generate contextually relevant responses [50].

The integration of KGs into LLMs significantly enhances reasoning and interaction capabilities, addressing critical limitations such as hallucinations and factual inaccuracies. This synergy improves model interpretation of complex graph data, resulting in increased accuracy and broader applicability across fields like education and automated question answering by providing structured, factual context that enhances reliability and performance [51, 43, 47, 52]. By leveraging the structured knowledge from KGs, LLMs can generate more sophisticated and contextually relevant outputs, paving the way for advanced AI systems.

4.2 Performance Optimization and Cost Management

Optimizing the performance and managing the costs of large language models (LLMs) require a multifaceted approach that balances computational demands with strategic deployment. Iterative Prompting Optimization (IPO) enhances model efficiency by allowing LLMs to refine solutions iteratively based on previous outputs and algorithmic instructions, minimizing loss functions and improving computational efficiency [53].

A significant challenge in deploying LLMs is the substantial computational resources required, including hardware and electricity costs [12]. Frameworks like LUNA provide a universal analysis platform to evaluate LLM quality from multiple trustworthiness perspectives, ensuring models are efficient and reliable [54]. Metrics designed to evaluate language modeling and downstream task performance are essential for a comprehensive view of model capabilities [55], guiding the development of strategies that enhance efficiency and effectiveness.

Probing techniques analyze LLM activations and predict outcomes based on internal representations, offering insights into performance optimization strategies [27]. These techniques facilitate a deeper understanding of model behavior, informing targeted improvements. The SES metric quantitatively measures the alignment of generated content with specified logical relations, providing a more accurate assessment of model reasoning [56].

The lack of standardized paradigms for evaluating LLM intelligence, particularly in distinguishing knowledge retrieval from problem-solving capabilities, remains a primary challenge [57]. Addressing this requires developing evaluation frameworks that accurately assess and enhance these distinct capabilities. Factor analysis techniques that empirically extract and understand the latent capabilities of LLMs represent a significant improvement, enabling targeted optimizations [58].

Libraries like PromptBench provide a flexible platform for evaluating LLMs, supporting original studies and comparisons across models while facilitating downstream application deployment [59]. By offering a standardized evaluation environment, PromptBench helps identify cost-effective strategies for LLM deployment, ensuring models are efficient and scalable.

The CONFIDENCE-DRIVEN INFERENCE method combines LLM annotations with confidence levels to optimize human annotation collection, aiming for accurate estimates with reduced human effort [60]. This approach emphasizes the importance of integrating human feedback with model outputs to enhance accuracy and efficiency.

These strategies highlight the critical role of innovative methodologies and frameworks in optimizing LLM performance while addressing the financial implications of their implementation. They underscore the necessity for models to be generalizable across applications, evaluable through established performance metrics, and deployed cost-effectively. As illustrated in Figure 4, which depicts the strategies for optimizing performance and managing costs in large language models, the figure highlights key optimization techniques, evaluation frameworks, and cost management strategies. By proposing tailored frameworks for generalization, evaluation, and cost modeling, this research elucidates the complexities involved in LLM development and management, as well as the importance of data augmentation techniques to enhance model capabilities without incurring additional data collection costs. Moreover, exploring retrieval-augmented methods illustrates how integrating external knowledge can improve the accessibility and comprehensibility of LLM-generated content, thereby broadening their applicability and effectiveness across diverse domains [61, 62, 6, 63, 64]. By adopting such approaches, the efficient and cost-effective utilization of LLMs across various applications becomes achievable.

5 Reinforcement Learning and Natural Language Processing

5.1 Challenges in Reinforcement Learning for NLP

Integrating reinforcement learning (RL) with natural language processing (NLP) presents significant challenges due to language's inherent ambiguity and the dynamic nature of RL environments [10]. A major challenge is the high computational demand of training deep networks, which requires extensive resources and complicates implementation [65]. The large action spaces typical in NLP tasks further complicate optimal action selection, increasing computational overhead.

Formulating effective reward functions is crucial for guiding RL agents in complex language tasks, yet defining clear state-action pairs remains difficult in many NLP applications [10]. This complexity hinders the training of robust models. Additionally, reliance on language models for supervision during distillation can propagate errors, leading to less robust smaller language models (SLMs)

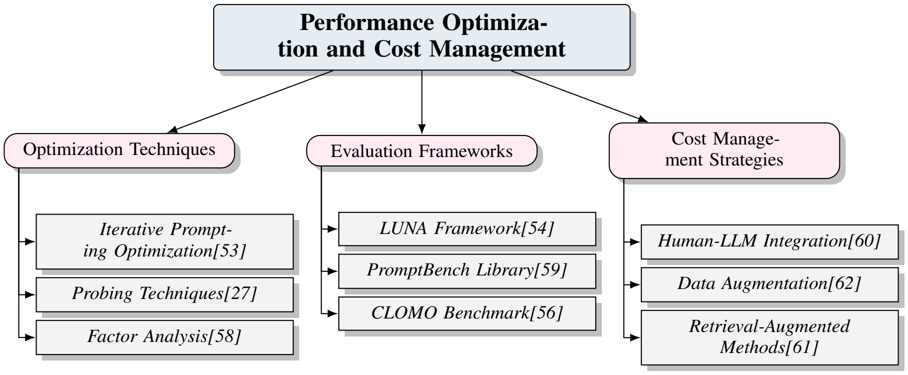


Figure 4: This figure illustrates the strategies for optimizing performance and managing costs in large language models, highlighting key optimization techniques, evaluation frameworks, and cost management strategies.

[60]. Miscalibrated confidence scores in large language models (LLMs) may result in suboptimal annotation selection, affecting RL-based systems' quality and reliability [60].

Existing benchmarks often fail to encompass the full spectrum of evaluation scenarios, especially regarding adversarial robustness and dynamic testing, limiting their real-world applicability [59]. Furthermore, accurately modeling moral choices in AI systems within social dilemmas highlights broader ethical issues in RL applications [66].

Comprehensive assessments of privacy risks and model capabilities in real-world scenarios are frequently overlooked in RL evaluations, representing a gap that needs to be addressed to develop robust and secure AI systems [67]. Addressing these challenges is essential for advancing RL integration in NLP, leading to more robust, interpretable, and effective AI systems capable of navigating complex language tasks.

5.2 Integration of Reinforcement Learning with LLMs

The integration of reinforcement learning (RL) with large language models (LLMs) marks a pivotal advancement in AI, enhancing decision-making and adaptive learning capabilities. Leveraging the extensive domain knowledge inherent in LLMs, RL facilitates intelligent modeling and strategic decision-making in dynamic environments. This synergy optimizes LLM architectures, improving output quality and enabling practical applications across various fields, including telecommunications and NLP. Such integration propels the potential for artificial general intelligence (AGI) while raising critical considerations regarding ethical alignment and human preferences in AI systems [10, 68, 69, 70, 71]. The adaptive learning mechanisms of RL combined with LLMs' proficiency in natural language understanding create sophisticated AI systems capable of nuanced reasoning and decision-making.

The MRKL system exemplifies this integration by merging LLMs with external knowledge sources and discrete reasoning components, demonstrating the potential of neuro-symbolic architectures to enhance reasoning capabilities [46]. This approach highlights the effectiveness of combining external knowledge with LLMs to improve performance in complex language tasks. GraphLLM, an end-to-end framework, further enhances LLM capabilities by incorporating graph learning models, enabling more effective reasoning about graph data [47]. This method showcases the application of RL techniques to refine LLMs, allowing them to adeptly process and reason with complex data structures.

RL integration with LLMs is enriched by frameworks that categorize NLP tasks as Markov decision processes (MDPs), optimizing decision-making processes within these tasks [10]. This categorization allows RL techniques to enhance LLM adaptability and performance in dynamic environments. Moreover, integrating multiple ethical theories, such as Utilitarianism, Deontological ethics, and Virtue Ethics, into RL frameworks emphasizes the necessity of defining moral reward structures within AI systems [66]. This ensures LLMs can navigate complex ethical landscapes, enhancing their reliability and trustworthiness across diverse applications.

AI-generated, for reference only.

The convergence of RL with LLMs establishes a powerful framework that amplifies AI capabilities, particularly in NLP tasks like conversational systems. Recent research illustrates how RL algorithms, especially those utilizing deep neural networks, can significantly enhance LLM performance and adaptability, such as ChatGPT, through methods like Reinforcement Learning from Human Feedback (RLHF). This integration addresses challenges in language understanding and generation, opening promising research avenues across domains like education and medicine [19, 10]. By harnessing the strengths of both RL and LLMs, these systems achieve more sophisticated, creative, and context-aware outputs, paving the way for innovative applications across diverse fields.

6 Human-AI Collaboration

6.1 Frameworks and Models for Human-AI Collaboration

Human-AI collaboration is enhanced through diverse frameworks and models that facilitate effective interaction. The Hybrid Reward Architecture (HRA) exemplifies this by employing multiple agents with decomposed reward functions to improve decision-making in complex environments [72]. The embodied cognition perspective underscores the importance of considering physical and contextual factors in AI design to align with human behaviors [73].

Integrating large language models (LLMs) with optimization algorithms enhances collaborative optimization processes [69], streamlining AI incorporation into human workflows. Trust dynamics are pivotal, necessitating models that analyze trust in human-AI contexts to develop reliable AI systems [14]. Middleware connecting UI affordances with LLM prompts exemplifies strategies to improve communication and collaboration [74].

Addressing biases in multimodal data is critical for equitable AI systems, with frameworks mitigating biases to support diverse user needs [75]. Multi-agent simulations reconceptualize language games, providing insights into emergent communication patterns for adaptive AI systems [76]. Benchmarks with modular designs enhance LLM capabilities in collaborative contexts, fostering robust AI systems [59].

These frameworks prioritize transparency, creativity, and safety, addressing LLM challenges and promoting a human-centered approach that enhances stakeholder understanding while mitigating deployment risks [11, 39, 1].

6.2 Challenges in Human-AI Collaboration

Integrating AI into human-centric environments presents challenges, notably bias in training data affecting fairness in decision-making, such as recruitment [75]. Ensuring AI systems do not perpetuate biases is essential for user trust. Variability in AI's ethical reasoning undermines reliability in moral decision-making [68], complicating the development of trustworthy systems aligned with human ethics.

Stealthy attacks that increase toxicity in LLM outputs threaten collaboration integrity, highlighting the need for robust security measures [77]. Ensuring reliability, transparency, and user comprehension of AI capabilities is crucial, as harmful automated decision mistakes necessitate transparent systems for informed user interaction [78].

Challenges in agent autonomy and communication in multi-agent reinforcement learning (MARL) underscore the complexity of designing effective AI systems [76]. Conflicting interests and rule observation gaps challenge safe AI deployment, necessitating regulatory frameworks for responsible use [39]. Interpretable decision aids that improve human performance in complex scenarios present opportunities to enhance collaboration [42].

Addressing these challenges ensures AI systems are reliable and ethical, supporting human activities across fields. This includes evaluating AI's role in processes like peer review, where AI boosts efficiency but raises ethical concerns about bias and transparency. Fostering transparency in LLM development aligns AI applications with stakeholder needs, facilitating responsible integration [11, 5].

6.3 Enhancing Trust and Transparency

Building trust and transparency in human-AI interactions is crucial for effective collaboration. Transparency involves clear communication about AI processes, enabling informed decisions, particularly in educational contexts where AI tutors support diverse learners, including those who are Deaf or Hard of Hearing (DHH) [33]. Frameworks must enable AI systems to articulate decision-making processes in comprehensible terms, considering diverse stakeholder needs [11, 13].

Trust is bolstered by designing AI systems with ethical considerations, including fairness and accountability. Robust evaluation frameworks assess AI performance across scenarios, identifying and addressing biases to promote equitable operation [79, 80, 1, 5, 6]. Comprehensive regulatory frameworks ensure compliance with ethical standards, addressing concerns related to bias, transparency, and accountability [5, 14].

A comprehensive strategy integrating clear communication, ethical design, and regulatory oversight enhances trust and transparency. This mitigates ethical concerns and fosters stakeholder understanding, ensuring responsible AI deployment across contexts [11, 13, 15, 14, 5]. Prioritizing these elements positions AI systems as reliable partners, enhancing human activities while maintaining user trust and confidence.

7 Ethical Considerations in AI

Ethical considerations are crucial in the development and deployment of AI technologies that significantly influence society. This section examines various dimensions of ethics in AI, particularly focusing on bias and fairness, which are deeply interconnected with societal values. Addressing bias in AI systems is vital for achieving equitable outcomes and ensuring that these technologies cater to the diverse needs of stakeholders. The subsequent subsection will explore the complexities of bias and fairness in AI, emphasizing the importance of comprehensive strategies to mitigate these challenges.

7.1 Bias and Fairness

Addressing bias and ensuring fairness in AI systems is essential due to their profound impact across various domains. Bias can originate from training data, algorithmic design, and deployment contexts. The opacity of decision-making processes in LLMs complicates bias identification and mitigation, necessitating robust interpretability mechanisms [58]. A lack of transparency may lead to misunderstandings about these systems' capabilities and limitations, potentially resulting in misuse or overreliance [81].

Incorporating human-centric approaches in AI design and deployment is crucial for equitable treatment across diverse demographic groups. Biases in LLMs, exacerbated by data scarcity for low-resource languages and imbalanced training data, hinder the integration of varied linguistic and cultural knowledge [82]. This underscores the necessity of aligning AI systems with local values and ethical standards, especially given the challenges in establishing a universally accepted ethical framework for AI agents in diverse social contexts [66].

The ethical implications of using LLMs in sensitive areas, such as mental health care, are significant. Concerns regarding deceptive empathy and the balance between helpfulness and harmlessness necessitate ongoing refinements to address biases and enhance applicability in real-world scenarios [56]. Vulnerabilities in LLMs, like those revealed through human feedback poisoning, further highlight the need for robust training methods and careful data collection practices to mitigate risks associated with bias and toxicity [67].

Integrating knowledge graphs can improve the fairness and reliability of AI outputs by incorporating type and neighborhood information, laying the groundwork for more equitable AI systems [83]. Interpretable decision aids are vital for addressing cognitive biases in human decision-making, providing insights into AI processes and fostering user trust [38]. However, gaps remain in understanding how fairness definitions vary across use cases and cultural contexts, necessitating further research to develop comprehensive frameworks for fairness certification in AI.

Recent advancements in steering the moral compass of LLMs toward various ethical schools demonstrate potential pathways to address biases within these models [82]. The introduction of novel reward functions based on the Value of Computation (VOC) also presents a promising approach to penalizing unnecessary reasoning, contributing to more balanced and fair AI systems [58].

Addressing bias and ensuring fairness in AI systems requires a multifaceted approach incorporating robust evaluation mechanisms, ethical design principles, and diverse perspectives. By prioritizing transparency, accountability, and fairness, AI systems can enhance their equity and trustworthiness, fostering ethical and responsible deployment across various domains, including scholarly peer review and information dissemination. This approach not only tackles common ethical challenges like biases and privacy concerns but also navigates the unique complexities associated with LLMs, such as hallucination and verifiable accountability. Implementing tailored ethical frameworks and dynamic auditing systems will further support the legitimacy of AI-driven processes, ensuring alignment with the moral and epistemic norms of their respective fields [5, 15].

7.2 Transparency and Accountability

Transparency and accountability in AI systems are critical for building trust and ensuring responsible deployment. Existing benchmarks often fall short in effectively communicating factuality and source attribution, essential for user understanding and trust in AI outputs [84]. The complexity, unpredictable behaviors, and proprietary nature of LLMs further complicate transparency efforts [11].

Accountability involves ensuring that AI systems operate within ethical and legal boundaries, including compliance with regulations like the GDPR [75]. This compliance is crucial for protecting user privacy and maintaining public trust in AI technologies. Jiao et al. emphasize the interconnected issues of bias, privacy, misinformation, and accountability, highlighting the need for comprehensive frameworks to address these challenges [15].

Transparency is particularly vital in crowdsourcing processes, where user engagement and clear communication about AI operations can enhance the effectiveness and reliability of collaborative efforts [3]. Furthermore, structured benchmarks that evaluate and compare the privacy protection capabilities of LLMs can identify critical areas for improvement, contributing to the development of more transparent and accountable AI systems [67].

To effectively enhance transparency and accountability, particularly in the context of LLMs, a comprehensive strategy is essential. This strategy should include clear communication tailored to diverse stakeholder needs, adherence to legal and ethical standards addressing unique challenges such as privacy, biases, and accountability, and the establishment of robust evaluation frameworks encompassing model reporting, evaluation results publication, and uncertainty communication. Integrating insights from human-computer interaction and interdisciplinary collaboration can create a more responsible and transparent AI ecosystem that mitigates risks and fosters trust among users and society [11, 5, 15]. By addressing these areas, AI systems can become more trustworthy and reliable, supporting ethical and responsible deployment across various domains.

7.3 Ethical Guidelines and Standards

Establishing ethical guidelines and standards for AI systems is imperative for ensuring their responsible and trustworthy deployment. The complexity of AI technologies, particularly in reinforcement learning from human feedback (RLHF) and partial observability, necessitates robust ethical frameworks that address potential risks and ensure accountability [39]. Integrating ethical considerations into the design and operation of human-AI interaction (HAI) systems is crucial for identifying and mitigating potential loopholes that could compromise safety and integrity [39].

The development of knowledge graph-enhanced large language models (KGLLMs) emphasizes the integration of knowledge graphs at various stages of model training, significantly enhancing the reasoning and interaction capabilities of AI systems [43]. This approach underscores the importance of incorporating structured knowledge into AI models to improve their ethical and functional performance.

The Guide-Align method proposes creating a comprehensive library of guidelines that can be dynamically matched to various inputs, thereby enhancing the safety and quality of LLM outputs [80]. This method highlights the necessity of adaptive ethical frameworks responsive to diverse contexts and inputs, ensuring that AI systems operate safely and effectively.

Privacy-preserving frameworks are critical for the ethical deployment of LLMs, protecting user data and maintaining public trust in AI technologies. The proposed benchmark by Ullah et al. aims to guide future research and development in privacy-aware AI, emphasizing the importance of privacy-security alignment in AI systems [85]. This aligns with the need for comprehensive privacy protection measures that safeguard user information while supporting AI functionality [67].

Future research should explore intricate moral frameworks and their implications for human-AI collaboration, particularly in modeling moral choices in social contexts [66]. This includes examining the impact of partner selection in moral agent interactions and developing more effective privacy-security alignment training methods [67].

Existing studies often lack comprehensive metrics for evaluating trust in AI systems, underscoring the need for robust evaluation frameworks addressing all identified trust dimensions [78]. Liao et al. organize current methods into four common approaches—model reporting, publishing evaluation results, providing explanations, and communicating uncertainty—each tailored to different stakeholder goals [11]. These methods are essential for enhancing transparency and accountability in AI systems.

The development and implementation of ethical guidelines and standards for AI require a multifaceted approach that incorporates diverse perspectives, robust evaluation mechanisms, and a commitment to privacy and safety. By emphasizing ethical frameworks, accountability mechanisms, and transparency in AI development and deployment, we can ensure that these technologies enhance efficiency and creativity while aligning with societal values and addressing critical challenges such as bias, privacy, and integrity in processes like peer review. This necessitates interdisciplinary collaboration and a human-centered perspective to navigate the complexities of AI ethics, ultimately fostering responsible AI integration that respects moral and epistemic norms within the scholarly community and beyond [1, 11, 5, 15].

7.4 Data Privacy and Security

Data privacy and security are pivotal in deploying and operating AI systems, particularly with the use of LLMs in sensitive domains. The rapid development of LLMs has intensified privacy concerns, especially with personal data potentially incorporated into training datasets [67]. This poses significant challenges in maintaining data confidentiality and integrity while ensuring the functionality and effectiveness of AI systems.

A primary challenge is the reliance on high-quality underlying models and the computational demands of implementing sophisticated methods, such as Bayesian models, which can affect the overall security and privacy of AI systems. Organizations deploying LLMs must manage technical debt carefully and evaluate the long-term implications of their deployment choices to maximize benefits while minimizing privacy risks. Considerations regarding LLMs must encompass adaptability to specific domain requirements while addressing the critical need for extensive data management, as improper handling could lead to significant privacy breaches. Additionally, the ethical implications surrounding LLMs necessitate the development of tailored frameworks and dynamic auditing systems to ensure accountability and transparency in their applications [86, 87, 40, 15, 67].

Constructing cross-data knowledge graphs, especially in languages with less robust NLP tool support, highlights privacy challenges related to clustering and intent discovery due to overlapping clusters. This necessitates the development of improved clustering techniques and privacy-preserving methods to ensure data integrity and confidentiality. Future work should focus on refining exposure scoring methodologies and exploring regulatory frameworks to mitigate the adverse effects of LLMs on labor markets, which can indirectly impact privacy and security [83].

Ethical considerations in AI research are crucial for upholding privacy standards, encompassing practices such as conducting experiments within controlled environments and ensuring the non-collection of private user information. These measures are essential not only to mitigate risks associated with biases and transparency issues in AI systems but also to align AI applications with broader moral and epistemic norms defining responsible conduct in the scholarly community. Moreover, the legitimacy of AI-driven processes, including peer review, depends on addressing the complex ethical challenges unique to these technologies, thereby fostering accountability and enhancing trust in their integration into various domains [5, 15]. Future research should focus on developing adaptable ethical frameworks that enhance transparency and accountability mechanisms, addressing societal impacts and privacy concerns associated with LLMs.

AI-generated, for reference only.

The use of synthetic data generated by AI systems presents both opportunities and challenges in mitigating privacy concerns. The reliability of synthetic data produced by LLMs and the unresolved privacy issues associated with their use in sensitive domains necessitate thorough investigation, particularly given the models' demonstrated inaccuracies in providing security and privacy advice and their potential to misinterpret or misrepresent expert knowledge [40, 87]. Enhancing safety filters and exploring the disinformation capabilities of LLMs in various languages are crucial steps toward improving data security.

Addressing data privacy and security in AI systems necessitates a multifaceted strategy that integrates advanced technical solutions—such as privacy-preserving mechanisms like differential privacy and robust ontology enrichment methods—with ethical considerations, including mitigating biases and accountability challenges inherent in LLMs. Continuous research is essential to adapt to evolving threats, as demonstrated by the need for dynamic auditing systems and interdisciplinary collaboration to enhance transparency and user trust. This comprehensive approach is crucial for effectively responding to the increasing complexities and vulnerabilities in the AI-driven technology landscape [87, 88, 15, 85, 41]. By prioritizing these elements, AI systems can be developed and deployed securely, ethically, and in alignment with societal values.

8 Conclusion

8.1 Future Directions and Ethical Considerations

Advancing research in hybrid human-AI systems necessitates a focus on optimizing multimodal models and investigating emerging trends in their applications to overcome existing limitations. Enhancing the emotional sensitivity and empathetic abilities of large language models (LLMs) is crucial for improving therapeutic interactions and care quality. The exploration of AI-Augmented Surveys utilizing LLMs for data collection and analysis must be accompanied by a careful consideration of ethical dimensions. Ensuring modality alignment and transparency in multimodal large language models (MLLMs) is vital for their safe application in sensitive domains such as healthcare. Developing comprehensive evaluation frameworks will facilitate the effective and ethical deployment of these systems, while also accommodating diverse user needs, including those of Deaf and Hard of Hearing (DHH) learners.

Refinement of ethical steering methods for LLMs is essential to align AI systems with varied human values and societal norms. Investigating the socio-technical implications of AI and establishing ethical alignment frameworks across different contexts remain critical. Additionally, enhancing prompt engineering techniques to improve the reliability of LLMs in providing security and privacy advice is a promising area for future inquiry. A nuanced intelligence evaluation framework that incorporates both quantitative and qualitative measures is necessary for a holistic assessment of AI capabilities, ensuring that evaluation metrics evolve alongside AI technologies. Moreover, improving metrics for long-text understanding will enhance LLM evaluation and applicability in complex real-world scenarios.

Innovative transparency strategies tailored to diverse stakeholders are crucial for building trust and accountability in AI systems. Future research should focus on establishing robust auditing mechanisms and examining the implications of LLM updates, aligning transparency efforts with stakeholder expectations and regulatory standards. Investigating the impact of the CREATIVITY INDEX in content creation and machine-generated text detection across various domains is also significant. Enhancing human-AI interaction processes within platforms like SciCrowd, improving data quality control, and supporting researchers in scientometric analyses are key areas for further exploration.

Furthermore, the development of specific metacognitive mechanisms for LLMs and their practical implementation in model training offers a promising research direction. Enhancing the calibration of LLM confidence scores and extending this approach to a broader range of NLP tasks could significantly bolster the reliability and applicability of LLMs. It is imperative that future research balances technological progress with ethical considerations, ensuring that hybrid human-AI systems are developed responsibly to enrich human experiences and contribute positively to societal outcomes.

References

[1] Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and Yejin Choi. Ai as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text, 2025.

[2] Wenxi Wang, Huansheng Ning, Feifei Shi, Sahraoui Dhelim, Weishan Zhang, and Liming Chen. A survey of hybrid human-artificial intelligence for social computing, 2021.

[3] A hybrid humanai tool for scien.

[4] Liming Chen. A survey of hybrid human-artificial intelligence for social computing.

[5] Laurie A. Schintler, Connie L. McNeely, and James Witte. A critical examination of the ethics of ai-mediated peer review, 2023.

[6] Jonas Becker, Jan Philip Wahle, Bela Gipp, and Terry Ruas. Text generation: A systematic literature review of tasks, evaluation, and challenges, 2024.

[7] Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a q-function, 2024.

[8] Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. Is neuro-symbolic ai meeting its promise in natural language processing? a structured review, 2022.

[9] Muhan Lin, Shuyang Shi, Yue Guo, Behdad Chalaki, Vaishnav Tadiparthi, Ehsan Moradi Pari, Simon Stepputtis, Joseph Campbell, and Katia Sycara. Navigating noisy feedback: Enhancing reinforcement learning with error-prone language models, 2024.

[10] Victor Uc-Cetina, Nicolas Navarro-Guerrero, Anabel Martin-Gonzalez, Cornelius Weber, and Stefan Wermter. Survey on reinforcement learning for language processing, 2022.

[11] Q. Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap, 2023.

[12] Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, Michael Gienger, Mathias Franzius, and Julian Eggert. A glimpse in chatgpt capabilities and its impact for ai research, 2023.

[13] Subhankar Maity and Aniket Deroy. Human-centric explainable ai in education, 2024.

[14] Andrea Ferrario, Michele Loi, and Eleonora Viganò. In ai we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology*, 33(3):523–539, 2020.

[15] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. Navigating llm ethics: Advancements, challenges, and future directions, 2024.

[16] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine, 2024.

[17] Mandana Saebi, Steven Krieg, Chuxu Zhang, Meng Jiang, and Nitesh Chawla. Heterogeneous relational reasoning in knowledge graphs with reinforcement learning, 2020.

[18] Ahmed Akakzia, Cédric Colas, Pierre-Yves Oudeyer, Mohamed Chetouani, and Olivier Sigaud. Grounding language to autonomously-acquired skills via goal generation, 2021.

[19] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of chatgpt-related research and perspective towards the future of large language models, 2023.

[20] Adarsh N L, Arun P V au2, and Aravindh N L. Enhancing image caption generation using reinforcement learning with human feedback, 2024.

[21] Karim Benharrak, Tim Zindulka, and Daniel Buschek. Deceptive patterns of intelligent and interactive writing assistants, 2024.

[22] Leon Lang, Davis Foote, Stuart Russell, Anca Dragan, Erik Jenner, and Scott Emmons. When your ais deceive you: Challenges of partial observability in reinforcement learning from human feedback, 2024.

[23] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Jiachen Li, Jennifer Bagdasarian, Guodong Gao, and Dakuo Wang. "i wish there were an ai": Challenges and ai potential in cancer patient-provider communication, 2024.

[24] Emile van Krieken. Optimisation in neurosymbolic learning systems, 2024.

[25] Bernard J. Koch and David Peterson. From protoscience to epistemic monoculture: How benchmarking set the stage for the deep learning revolution, 2024.

[26] Eleni Nisioti and Clément Moulin-Frier. Grounding artificial intelligence in the origins of human behavior, 2020.

[27] Simon Goldstein and Benjamin A. Levinstein. Does chatgpt have a mind?, 2024.

[28] Aparna Elangovan, Ling Liu, Lei Xu, Sravan Bodapati, and Dan Roth. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models, 2024.

[29] Zainab Iftikhar, Sean Ransom, Amy Xiao, and Jeff Huang. Therapy as an nlp task: Psychologists' comparison of llms and human peers in cbt, 2024.

[30] Mihael Arcan, David-Paul Niland, and Fionn Delahunty. An assessment on comprehending mental health through large language models, 2024.

[31] Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Junyu Liu, Benji Peng, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. From text to multimodality: Exploring the evolution and impact of large language models in medical practice, 2024.

[32] Yingzhe Peng, Xiaoting Qin, Zhiyang Zhang, Jue Zhang, Qingwei Lin, Xu Yang, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. Navigating the unknown: A chat-based collaborative interface for personalized exploratory tasks, 2024.

[33] Haocong Cheng, Si Chen, Christopher Perdriau, and Yun Huang. Llm-powered ai tutors with personas for d/deaf and hard-of-hearing online learners, 2024.

[34] Jeremy Straub and Zach Johnson. Initial development and evaluation of the creative artificial intelligence through recurring developments and determinations (cairdd) system, 2024.

[35] Zilong Wang, Nan Chen, Luna K. Qiu, Ling Yue, Geli Guo, Yang Ou, Shiqi Jiang, Yuqing Yang, and Lili Qiu. The potential and value of ai chatbot in personalized cognitive training, 2024.

[36] Roma Shusterman, Allison C. Waters, Shannon O'Neill, Phan Luu, and Don M. Tucker. An active inference strategy for prompting reliable responses from large language models in medical practice, 2024.

[37] Jaime F. Fisac, Monica A. Gates, Jessica B. Hamrick, Chang Liu, Dylan Hadfield-Menell, Malayandi Palaniappan, Dhruv Malik, S. Shankar Sastry, Thomas L. Griffiths, and Anca D. Dragan. Pragmatic-pedagogic value alignment, 2018.

[38] Self-critiquing models for assisting human evaluators.

[39] Stephen Fox and Juan G Victores. Safety of human–artificial intelligence systems: Applying safety science to analyze loopholes in interactions between human organizations, artificial intelligence, and individual people. In *Informatics*, volume 11, page 36. MDPI, 2024.

AI-generated, for reference only.

[40] Phoebe Jing, Yijing Gao, Yuanhang Zhang, and Xianlong Zeng. Translating expert intuition into quantifiable features: Encode investigator domain knowledge via llm for enhanced predictive analytics, 2024.

[41] Ken Gu, Ruoxi Shang, Tim Althoff, Chenglong Wang, and Steven M. Drucker. How do analysts understand and verify ai-assisted data analyses?, 2024.

[42] Julian Skirzyński, Frederic Becker, and Falk Lieder. Automatic discovery of interpretable planning strategies, 2021.

[43] Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling, 2024.

[44] Qingxiu Dong, Li Dong, Ke Xu, Guangyan Zhou, Yaru Hao, Zhifang Sui, and Furu Wei. Large language model for science: A study on p vs. np, 2023.

[45] Weize Liu, Guocong Li, Kai Zhang, Bang Du, Qiyuan Chen, Xuming Hu, Hongxia Xu, Jintai Chen, and Jian Wu. Mind's mirror: Distilling self-evaluation capability and comprehensive thinking from large language models, 2024.

[46] Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, Dor Muhlgay, Noam Rozen, Erez Schwartz, Gal Shachaf, Shai Shalev-Shwartz, Amnon Shashua, and Moshe Tenenholtz. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning, 2022.

[47] Ziwei Chai, Tianjie Zhang, Liang Wu, Kaiqiao Han, Xiaohai Hu, Xuanwen Huang, and Yang Yang. Graphllm: Boosting graph reasoning ability of large language model, 2023.

[48] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. Recommender systems in the era of large language models (llms), 2024.

[49] Donghoon Shin, Gary Hsieh, and Young-Ho Kim. Planfitting: Tailoring personalized exercise plans with large language models, 2023.

[50] Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios, 2024.

[51] Tuan Bui, Oanh Tran, Phuong Nguyen, Bao Ho, Long Nguyen, Thang Bui, and Tho Quan. Cross-data knowledge graph construction for llm-enabled educational question-answering system: A case study at hcmut, 2024.

[52] Ernests Lavrinovics, Russa Biswas, Johannes Bjerva, and Katja Hose. Knowledge graphs, large language models, and hallucinations: An nlp perspective, 2024.

[53] Pei-Fu Guo, Ying-Hsuan Chen, Yun-Da Tsai, and Shou-De Lin. Towards optimizing with large language models, 2024.

[54] Da Song, Xuan Xie, Jiayang Song, Derui Zhu, Yuheng Huang, Felix Juefei-Xu, and Lei Ma. Luna: A model-based universal analysis framework for large language models, 2024.

[55] Yutong Hu, Quzhe Huang, Mingxu Tao, Chen Zhang, and Yansong Feng. Can perplexity reflect large language model's ability in long text understanding?, 2024.

[56] Yinya Huang, Ruixin Hong, Hongming Zhang, Wei Shao, Zhicheng Yang, Dong Yu, Changshui Zhang, Xiaodan Liang, and Linqi Song. Clomo: Counterfactual logical modification with large language models, 2024.

[57] Nils Körber, Silvan Wehrli, and Christopher Irrgang. How to measure the intelligence of large language models?, 2024.

[58] Ryan Burnell, Han Hao, Andrew R. A. Conway, and Jose Hernandez Orallo. Revealing the structure of language model capabilities, 2023.

[59] Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. Promptbench: A unified library for evaluation of large language models, 2024.

[60] Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candès, and Dan Jurafsky. Can unconfident llm annotations be used for confident conclusions?, 2025.

[61] Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. Retrieval augmentation of large language models for lay language generation, 2024.

[62] Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. Data augmentation using large language models: Data perspectives, learning paradigms and challenges, 2024.

[63] Abi Aryan, Aakash Kumar Nain, Andrew McMahon, Lucas Augusto Meyer, and Harpreet Singh Sahota. The costly dilemma: Generalization, evaluation and cost-optimal deployment of large language models, 2023.

[64] Lochan Basyal and Mihir Sanghvi. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models, 2023.

[65] Sarvesh Patil. Deep learning based natural language processing for end to end speech translation, 2018.

[66] Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Modeling moral choices in social dilemmas with multi-agent reinforcement learning, 2023.

[67] Yuqi Yang, Xiaowen Huang, and Jitao Sang. Exploring the privacy protection capabilities of chinese large language models, 2024.

[68] Alejandro Tlaie. Exploring and steering the moral compass of large language models, 2024.

[69] Sen Huang, Kaixiang Yang, Sheng Qi, and Rui Wang. When large language model meets optimization, 2024.

[70] Hao Zhou, Chengming Hu, Ye Yuan, Yufei Cui, Yili Jin, Can Chen, Haolun Wu, Dun Yuan, Li Jiang, Di Wu, Xue Liu, Charlie Zhang, Xianbin Wang, and Jiangchuan Liu. Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities, 2024.

[71] Adam X. Yang, Maxime Robeyns, Thomas Coste, Zhengyan Shi, Jun Wang, Haitham Bou-Ammar, and Laurence Aitchison. Bayesian reward models for llm alignment, 2024.

[72] Harm van Seijen, Mehdi Fatemi, Joshua Romoff, Romain Laroche, Tavian Barnes, and Jeffrey Tsang. Hybrid reward architecture for reinforcement learning, 2017.

[73] Sara Incao, Carlo Mazzola, Giulia Belgiovine, and Alessandra Sciutti. A roadmap for embodied and social grounding in llms, 2024.

[74] Stephen MacNeil, Andrew Tran, Joanne Kim, Ziheng Huang, Seth Bernstein, and Dan Mogil. Prompt middleware: Mapping prompts for large language models to ui affordances, 2023.

[75] Alejandro Peña, Ignacio Serna, Aythami Morales, Julian Fierrez, Alfonso Ortega, Ainhoa Herrarte, Manuel Alcantara, and Javier Ortega-Garcia. Human-centric multimodal machine learning: Recent advances and testbed on ai-based recruitment, 2023.

[76] Paul Van Eecke and Katrien Beuls. Re-conceptualising the language game paradigm in the framework of multi-agent reinforcement learning, 2020.

[77] Bocheng Chen, Hanqing Guo, Guangjing Wang, Yuanda Wang, and Qiben Yan. The dark side of human feedback: Poisoning large language models via user inputs, 2024.

[78] Sivan Schwartz, Avi Yaeli, and Segev Shlomov. Enhancing trust in llm-based ai automation agents: New considerations and future challenges, 2023.

[79] Suriya Prakash Jambunathan, Ashwath Shankarnarayan, and Parijat Dube. Convnlp: Image-based ai text detection, 2024.

[80] Yi Luo, Zhenghao Lin, Yuhao Zhang, Jiashuo Sun, Chen Lin, Chengjin Xu, Xiangdong Su, Yelong Shen, Jian Guo, and Yeyun Gong. Ensuring safe and high-quality outputs: A guideline library approach for language models, 2024.

[81] Florian Scholten, Tobias R. Rebholz, and Mandy Hütter. Metacognitive myopia in large language models, 2024.

[82] Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. A survey on human preference learning for large language models, 2024.

[83] Qin Chen, Jinfeng Ge, Huaqing Xie, Xingcheng Xu, and Yanqing Yang. Large language models at work in china's labor market, 2023.

[84] Hyo Jin Do, Rachel Ostrand, Justin D. Weisz, Casey Dugan, Prasanna Sattigeri, Dennis Wei, Keerthiram Murugesan, and Werner Geyer. Facilitating human-llm collaboration through factuality scores and source attributions, 2024.

[85] Imdad Ullah, Najm Hassan, Sukhpal Singh Gill, Basem Suleiman, Tariq Ahamed Ahanger, Zawar Shah, Junaid Qadir, and Salil S. Kanhere. Privacy preserving large language models: Chatgpt case study based vision and framework, 2023.

[86] Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. Disinformation capabilities of large language models, 2024.

[87] Yufan Chen, Arjun Arunasalam, and Z. Berkay Celik. Can large language models provide security privacy advice? measuring the ability of llms to refute misconceptions, 2023.

[88] Lalit Mohan Sanagavarapu, Vivek Iyer, and Raghu Reddy. A deep learning approach for ontology enrichment from unstructured text, 2021.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.