In AAAI 2025 Spring Symposium on Current and Future Varieties of Human-AI Collaboration (HAI 2025)
San Francisco, CA, USA

Leveraging LLMs for Collaborative Human-AI Decision Making

Anthony Favier, Pulkit Verma, Ngoc La, and **Julie A. Shah**

MIT AeroAstro, Cambridge, USA
{antfav24, pulkitv, ntmla}@mit.edu, julie_a_shah@csail.mit.edu

Abstract

Human-AI collaboration is a rapidly evolving field that seeks to leverage the complementary strengths of humans and artificial intelligence (AI) to solve complex problems. An area where such collaboration holds significant promise is in decision-making tasks, particularly in automated planning. As classical symbolic approaches are widely used in this field, they are limited when solving large and complex problems. Furthermore, they require expert knowledge in formal and structured languages to interact with, hindering their use. Recently, Large Language Models (LLMs) have emerged as a potential solution to these challenges but LLMs alone are not sufficient for solving such problems. However, a promising way to achieve seamless human-AI collaboration could be with hybrid approaches combining the strength of symbolic reasoning and the flexibility of LLMs.

Introduction

Human-AI collaboration is a rapidly evolving field that seeks to leverage the complementary strengths of humans and artificial intelligence (AI) to solve complex problems. An area where such collaboration holds significant promise is in decision-making tasks, particularly in automated planning (Ghallab, Nau, and Traverso 2016). Classical automated planners have been widely used to address such problems, but they often struggle with large, complex, and realistic scenarios. Due to their PSPACE complexity (Bylander 1991), these systems typically cannot generate optimal solutions for complex problems. Instead, they rely on simplification techniques like heuristics and relaxation methods to manage computational demands (Müller-Merbach 1981). While effective in constrained settings, these approaches often fall short when applied to real-world problems that require adaptability and scalability.

In recent years, Large Language Models (LLMs) have emerged as a potential solution to these challenges. LLMs, with their ability to process and generate natural language, offer a unique combination of abstraction and generalization. However, despite their promise, LLMs alone are not sufficient to solve complex planning problems (Kambhampati et al. 2024). Their limitations in reasoning, consistency, and reliability highlight the need for hybrid approaches that combine the strengths of symbolic reasoning with the flexibility of LLMs. An overview of such hybrid approach is shown in Figure 1. In this extended abstract, we use a disaster recovery scenario as a running example to discuss how to leverage LLMs in a Human-AI collaboration.

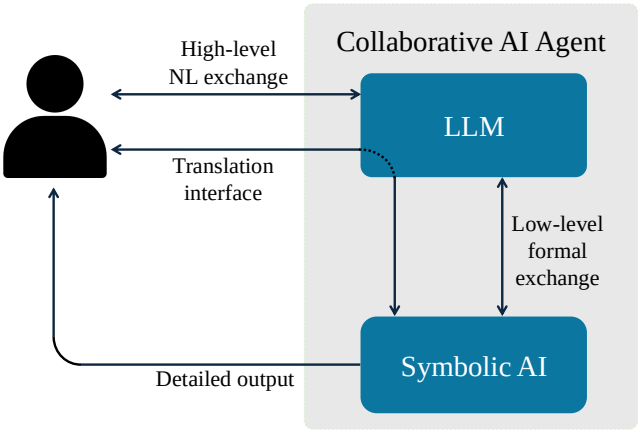

Figure 1: Overview of a generic Hybrid Collaborative AI Agent

Human-AI Collaboration Roles in Decision-Making

In this context, different roles can be identified. Classical symbolic AI should be in charge of *low-level, complex and long-horizon reasoning*. They ensure sound and detailed solutions while considering formal constraints. In the disaster recovery scenario, they can calculate and compare complete rescue plans. However, despite being mandatory for computing full sound solutions, classical AI lacks high-level contextual considerations and can be computationally expensive.

On the other hand, humans should act as *high-level reasoners and critics*. Their inherent common sense and intuition allow for estimating risks and identifying promising strategies that align, for instance, with ethical and practical considerations. Since humans struggle with long and low-level calculations, they leverage symbolic systems to handle these calculations, but expert knowledge in formal planning and programming languages is required, limiting the acces-

sibility of such tools.

In this context, LLMs can play several roles beneficial to symbolic systems, humans, and the collaboration. First, LLMs can act as an *interface translator* converting human natural language inputs into formal, structured language (e.g. PDDL (McDermott et al. 1998)) for symbolic systems. This would avoid the need for specialized technical knowledge, which can impede rapid response and effective coordination. For instance, an LLM could translate a human operator's verbal instruction to "prioritize children and injured individuals" into a formal constraint for the planner. LLMs can *guide the search process* of symbolic planners by informing heuristic functions or suggesting high-level strategies. For instance, in a disaster scenario, an LLM might propose prioritizing rescue routes with lower risks or higher practicability based on weather predictions and fuel constraints. LLMs can *generate ideas* to explore novel alternative solutions that may not be immediately apparent to human planners. For example, an LLM might suggest using drones to scout inaccessible areas or rerouting ground vehicles to avoid flooded paths. LLMs can also *highlight information* to emphasize critical information within a problem domain, such as areas with high survival probabilities or locations with rapidly deteriorating weather conditions.

LLM Challenges and Considerations While LLMs offer promising capabilities for natural language processing, recent benchmarking studies reveal significant limitations in their planning and reasoning abilities. They can hallucinate, misinterpret, or omit information, leading to errors. Even their translation capabilities from formal syntax to natural language has been shown to be inaccurate (Karia et al. 2024; Parmar et al. 2024). Kokel et al. (2025) introduced ACPBench, a comprehensive benchmark for reasoning about action, change, and planning that evaluates LLM performance across multiple dimensions. Their findings, comparing various state-of-the-art models (including Phi-3, Gemma 7B, LLAMA 3.1, GPT-4o, and others), demonstrate that even the most advanced LLMs struggle with fundamental aspects of planning. The benchmark evaluates models on metrics including atom reachability, progression, validation, action applicability, landmarks, justification, and action reachability. Results show that while LLMs perform moderately well on some dimensions, they exhibit substantial weaknesses in others, with no model achieving strong performance across all metrics. This empirical evidence further supports Kambhampati et al. (2024)'s analysis that LLMs can be used for planning, but not for doing end-to-end planning, reinforcing our argument for hybrid approaches that combine LLMs with symbolic reasoning systems rather than relying on LLMs alone for complex decision-making tasks.

Incorporating Transparency, Trust, and Dependability

Effective Human-AI collaboration requires transparency, trust, and dependability. Transparency helps operators understand AI decisions, such as why a particular route was prioritized or why a rescue team was assigned to a specific location. Liao and Vaughan (Liao and Wortman Vaughan 2024) suggest transparency approaches of providing model reports, publishing evaluation results, communicating uncertainty, and including explanations. On the other hand, Ehsan et al. (Ehsan et al. 2024) argue that algorithmic transparency is not enough for making AI explainable. LLM transparency should be a human-centered, socio-technical interaction design problem, in which supporting humans make sense of the model is more critical than peeking into its architecture. These various aspects of the transparency problem emphasize the need to specify what information human users need while working with LLMs or any AI tools.

Dependability requires robust mechanisms to handle errors and uncertainties, especially when LLMs are involved. Uncertainties or ambiguous outputs can be flagged for human review, ensuring that critical decisions are not based on unreliable information. For example, LLM-generated routes can be cross-checked with weather data and human input. Kim et al. (Kim et al. 2024) introduce Prometheus 2, a powerful language model designed to evaluate other LLMs, thereby contributing to improved dependability. Jung et al. (Jung, Brahman, and Choi 2025) propose a cascaded selective evaluation approach, which dynamically switches between weaker and stronger LLMs to align better with human expectations. However, a key unresolved question remains: what happens when even the most capable LLMs still fail to meet human standards?

Trust is built through consistent and reliable performance, especially in high-stakes situations where errors can have severe consequences. Transparency and dependability are key factors in earning this trust. While much research has focused on enhancing trust in AI systems, excessive trust can also be problematic. When humans overestimate an AI system's capabilities, they risk being misled by inaccurate outputs, potentially undermining effective collaboration (Musaffar et al. 2025). Therefore, trust must be carefully calibrated in designing LLM-based interaction interfaces.

Measuring Performance and Addressing Tradeoffs

In measuring performance within Human-AI collaboration, it's essential to evaluate both the individual contributions of the LLM-based AI tools and the human users, as well as the combined system as a whole. On the AI side, different roles such as planning, translation, and suggestion, require tailored evaluation metrics. These may include metrics like success rate and reward for planning tasks (Huang et al. 2024), consistency, plausibility, and stability for translation tasks (Perko and Wotawa 2024), and accuracy, diversity, and fairness for suggestion tasks (Gao et al. 2024). Such tailored metrics enable a more precise evaluation of LLM-based tools' effectiveness in fulfilling their intended roles.

On the human side, cognitive metrics are assessed through user studies, with a focus on trust and adaptability. Tools such as the Trust of Automated Systems Test (TOAST) and TrustDiff are used to quantitatively assess these metrics both before and after user interaction with AI (Oelschlager 2024). These metrics help illuminate the evolving dynamics of human reliance on and collaboration with AI systems.

Beyond evaluating individual contributions, it is essential to assess the overall Human-AI system as a combined entity. This includes a mix of qualitative metrics, such as solution quality, human satisfaction, and cognitive load, and quantitative metrics, such as task completion time, planning efficiency, and usability of the system. In addition, benefit-risk tradeoffs must be considered, encompassing ethical concerns, risk management, task performance, and the need for human oversight in critical decision-making scenarios. For example, while LLMs can improve decision-making by providing insights and reducing cognitive load, they may also introduce biases, such as prioritizing survivors with higher probabilities, which could raise ethical concerns. Balancing these tradeoffs demands careful system design and continuous evaluation. While evaluating each component individually provides valuable details, a holistic approach offers a comprehensive understanding of the overall effectiveness and impact of Human-AI collaboration.

Conclusion

The integration of LLMs into collaborative decision-making systems represents a promising avenue for addressing complex automated planning problems. By combining the strengths of symbolic reasoning and LLMs, hybrid approaches can overcome the limitations of traditional methods while enabling more accessible and effective human-AI collaboration. However, realizing this potential requires addressing challenges related to verification, transparency, and dependability, as well as developing robust performance metrics and strategies for managing benefit-risk tradeoffs. As research in this area continues to advance, it has the potential to transform how humans and AI systems work together to solve some of the most challenging problems across diverse application domains.

Acknowledgments

This work was supported by the ONR under grant N000142312883.

References

Bylander, T. 1991. Complexity Results for Planning. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI)*.

Ehsan, U.; Watkins, E.; Wintersberger, P.; Manger, C.; Kim, S.; Van Berkel, N.; Riener, A.; and Riedl, M. 2024. Human-Centered Explainable AI (HCXAI): Reloading Explainability in the Era of Large Language Models (LLMs). In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Sytems (CHI)*.

Gao, J.; Chen, B.; Zhao, X.; Liu, W.; Li, X.; Wang, Y.; Zhang, Z.; Wang, W.; Ye, Y.; Lin, S.; et al. 2024. LLM4Rerank: LLM-based Auto-Reranking Framework for Recommendations. *arXiv preprint arXiv:2406.12433*.

Ghallab, M.; Nau, D.; and Traverso, P. 2016. *Automated Planning and Acting*. Cambridge University Press.

Huang, X.; Liu, W.; Chen, X.; Wang, X.; Wang, H.; Lian, D.; Wang, Y.; Tang, R.; and Chen, E. 2024. Understanding the Planning of LLM Agents: A Survey. *arXiv preprint arXiv:2402.02716*.

Jung, J.; Brahman, F.; and Choi, Y. 2025. Trust or Escalate: LLM Judges with Provable Guarantees for Human Agreement. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Kambhampati, S.; Valmeekam, K.; Guan, L.; Verma, M.; Stechly, K.; Bhambri, S.; Saldyt, L. P.; and Murthy, A. B. 2024. Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Karia, R.; Dobhal, D.; Bramblett, D.; Verma, P.; and Srivastava, S. 2024. ∀uto∃val: Autonomous Assessment of LLMs in Formal Synthesis and Interpretation Tasks. arXiv:2403.18327.

Kim, S.; Suk, J.; Longpre, S.; Lin, B. Y.; Shin, J.; Welleck, S.; Neubig, G.; Lee, M.; Lee, K.; and Seo, M. 2024. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kokel, H.; Katz, M.; Srinivas, K.; and Sohrabi, S. 2025. ACPBench: Reasoning about Action, Change, and Planning. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*.

Liao, Q. V.; and Wortman Vaughan, J. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, (Special Issue 5).

McDermott, D.; Ghallab, M.; Howe, A.; Knoblock, C.; Ram, A.; Veloso, M.; Weld, D. S.; and Wilkins, D. 1998. PDDL – The Planning Domain Definition Language. Technical report, Yale Center for Computational Vision and Control.

Musaffar, A. K.; Gokhale, A.; Zeng, S.; Tadayon, R.; Yan, X.; Singh, A.; and Bullo, F. 2025. Learning to Lie: Reinforcement Learning Attacks Damage Human-AI Teams and Teams of LLMs. In *ICLR 2025 Workshop on Human-AI Coevolution*.

Müller-Merbach, H. 1981. Heuristics and Their Design: A Survey. *European Journal of Operational Research*, 8(1): 1–23.

Oelschlager, R. 2024. *Evaluating the Impact of Hallucinations on User Trust and Satisfaction in LLM-based Systems*. Bachelor's thesis, Linnaeus University, Department of Computer Science and Media Technology.

Parmar, M.; Varshney, N.; Patel, N.; Luo, M.; Mashetty, S.; Mitra, A.; and Baral, C. 2024. LogicBench: Towards Systematic Evaluation of Logical Reasoning Ability of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Perko, A.; and Wotawa, F. 2024. Evaluating OpenAI Large Language Models for Generating Logical Abstractions of Technical Requirements Documents. In *2024 IEEE 24th International Conference on Software Quality, Reliability and Security (QRS)*, 238–249.