Massachusetts Institute of Technology

2025 Spring Symposium Series
H-AI Symposium

Association for the Advancement of Artificial Intelligence

Leveraging LLMs for Collaborative Human-AI Decision Making

March 31 2025

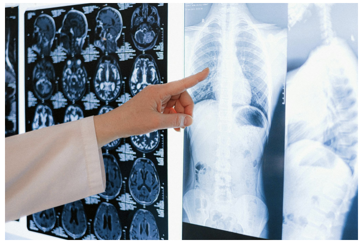
Anthony Favier, Pulkit Verma, Ngoc La, Julie A. Shah

Interactive Robotics Group
AeroAstro and CSAIL

Introduction

Human-AI Collaboration is a rapidly **evolving** and **promising** field

Healthcare



***Accelerates** the **diagnostic** process and **enhances** the **accuracy** of results.*

Finance

*Provides **early warnings** about potential market shifts or emerging opportunities.*

Art

*Offers **suggestions** and **refining** the output based on user **preferences**.*

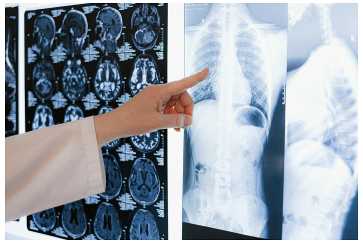
Manufacturing

***Collaborative robots** handle the physically demanding, repetitive tasks, **reducing** human **fatigue** and **injuries**.*

Introduction

Human-AI Collaboration is a rapidly **evolving** and **promising** field

Healthcare



***Accelerates** the **diagnostic** process and **enhances** the **accuracy** of results.*

Finance

*Provides **early warnings** about potential market shifts or emerging opportunities.*

Art

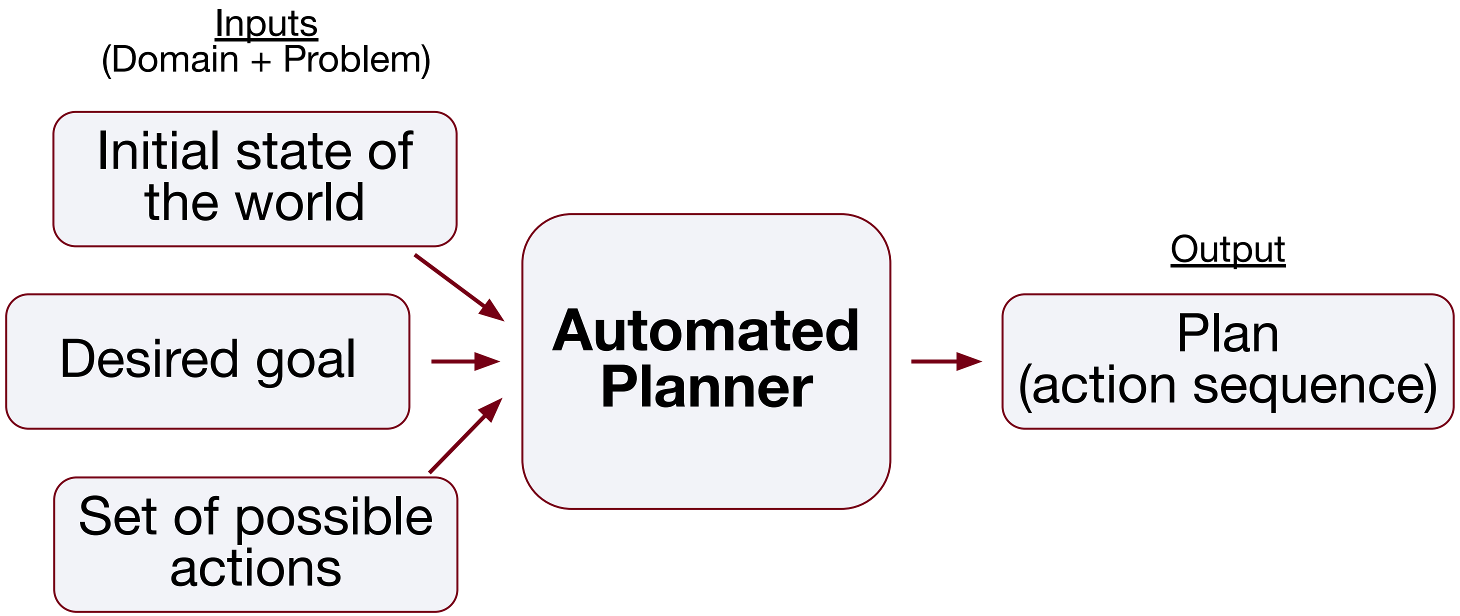
*Offers **suggestions** and **refining** the output based on user **preferences**.*

Manufacturing

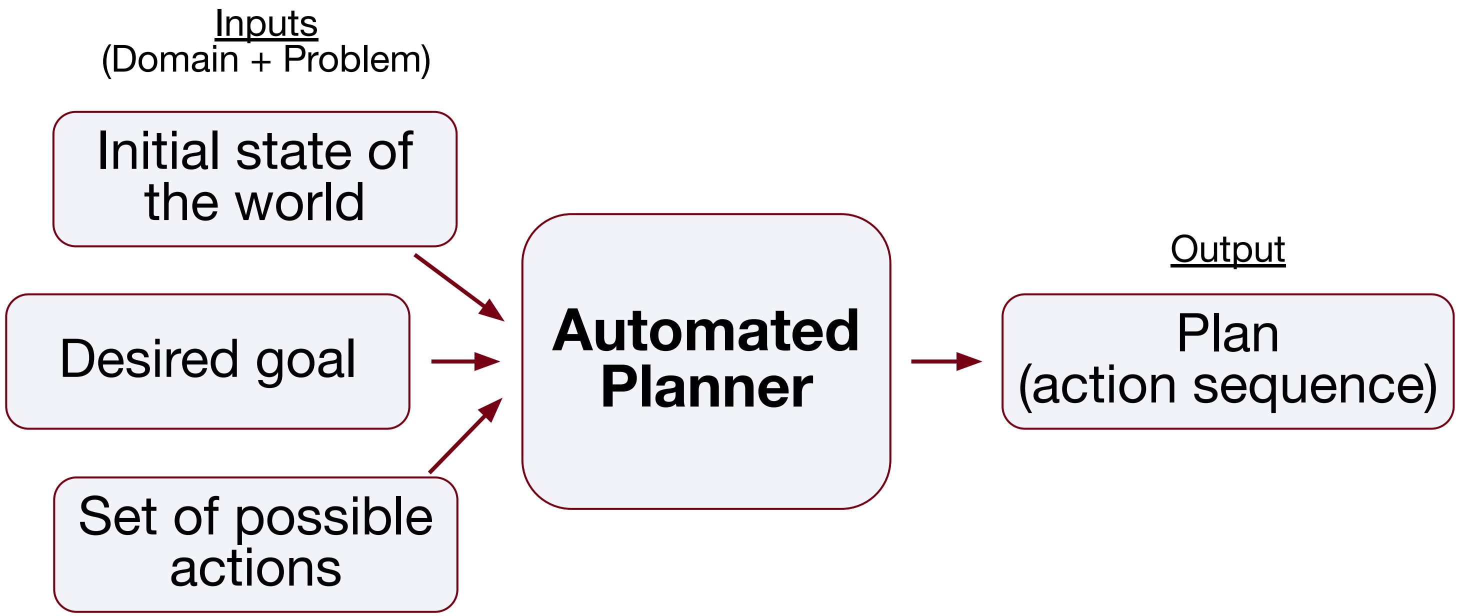
***Collaborative robots** handle the physically demanding, repetitive tasks, **reducing** human **fatigue** and **injuries**.*

AI **can't** replace the **human expertise** but rather **enhances** their capabilities, creating a **synergy** that improves **productivity**.

Promising for Automated Planning



Promising for Automated Planning



Difficulty defined by:

- **Problem Size**
  - Number of objects
  - Number of actions
- **Simplifying assumptions**
  - Deterministic or not
  - Discrete or Continuous
  - Full or Partial Observability
  - Sequential or Concurrent
  - Single or Multi-Agent

Promising for Automated Planning

For **realistic** scenarios, automated planners **can't** generate **optimal** solutions.

Heuristics
&
Relaxation techniques → "**Good enough**" solutions

Not sufficient for high stakes scenarios

Promising for Automated Planning

For **realistic** scenarios, automated planners **can't** generate **optimal** solutions.

Heuristics
&
Relaxation techniques → **"Good enough"** solutions

Not sufficient for high stakes scenarios

Human high-level reasoning could **guide** the solving process.

However, **requires** expert knowledge in **formal programming and planning languages** (e.g. PDDL, Python, etc..)

LLMs as a potential solution

Their ability to **process** and **generate natural language** offers a unique combination of **abstraction** and **generalization**.

LLMs as a potential solution

Their ability to **process** and **generate natural language** offers a unique combination of **abstraction** and **generalization**.

However, LLMs alone are **not sufficient** for solving **complex planning problems** (Kambhampati et al., ICML 2024).

	Blocksworld	Obfuscated Blocksworld
SoTA LLMs	**60%**	**4%**

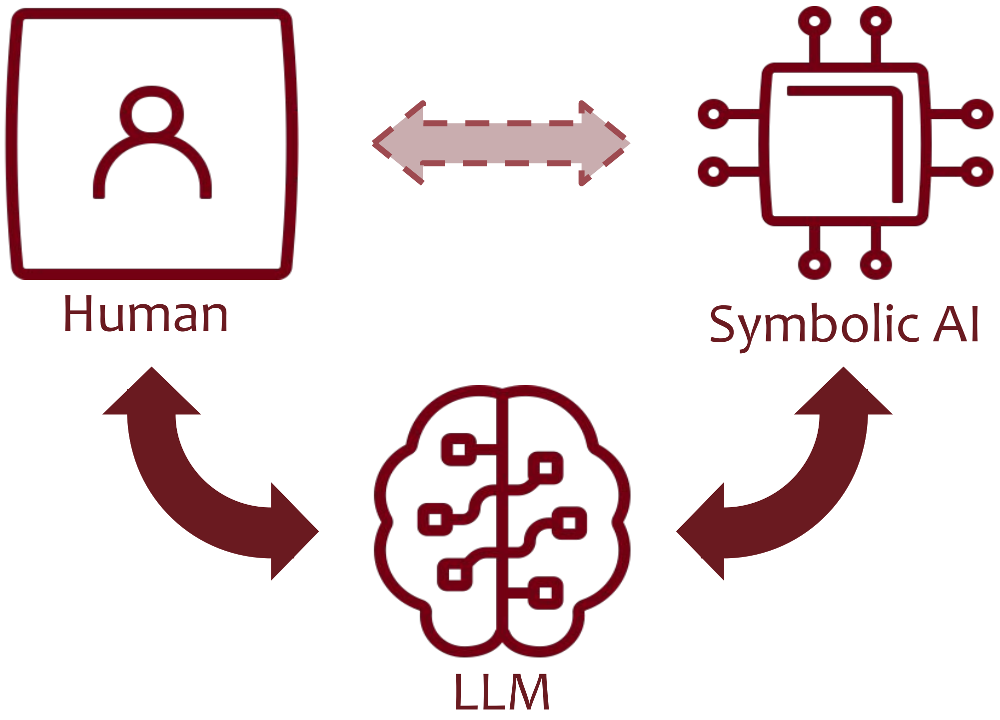
Success rates

They **suffer** from too many **limitations** in **reasoning**, **consistency** and **reliability**.

LLMs as a potential solution

This highlights the **need** for **hybrid approaches** combining the strengths of symbolic reasoning with the flexibility of LLMs.

More precisely, considering the **natural language processing capabilities** of LLMs, we believe that **hybrid human-in-the-loop approaches** are very **promising**.



Running Example: Disaster Recovery

Goal: **Rescue** groups of **persons** endangered by a natural disaster

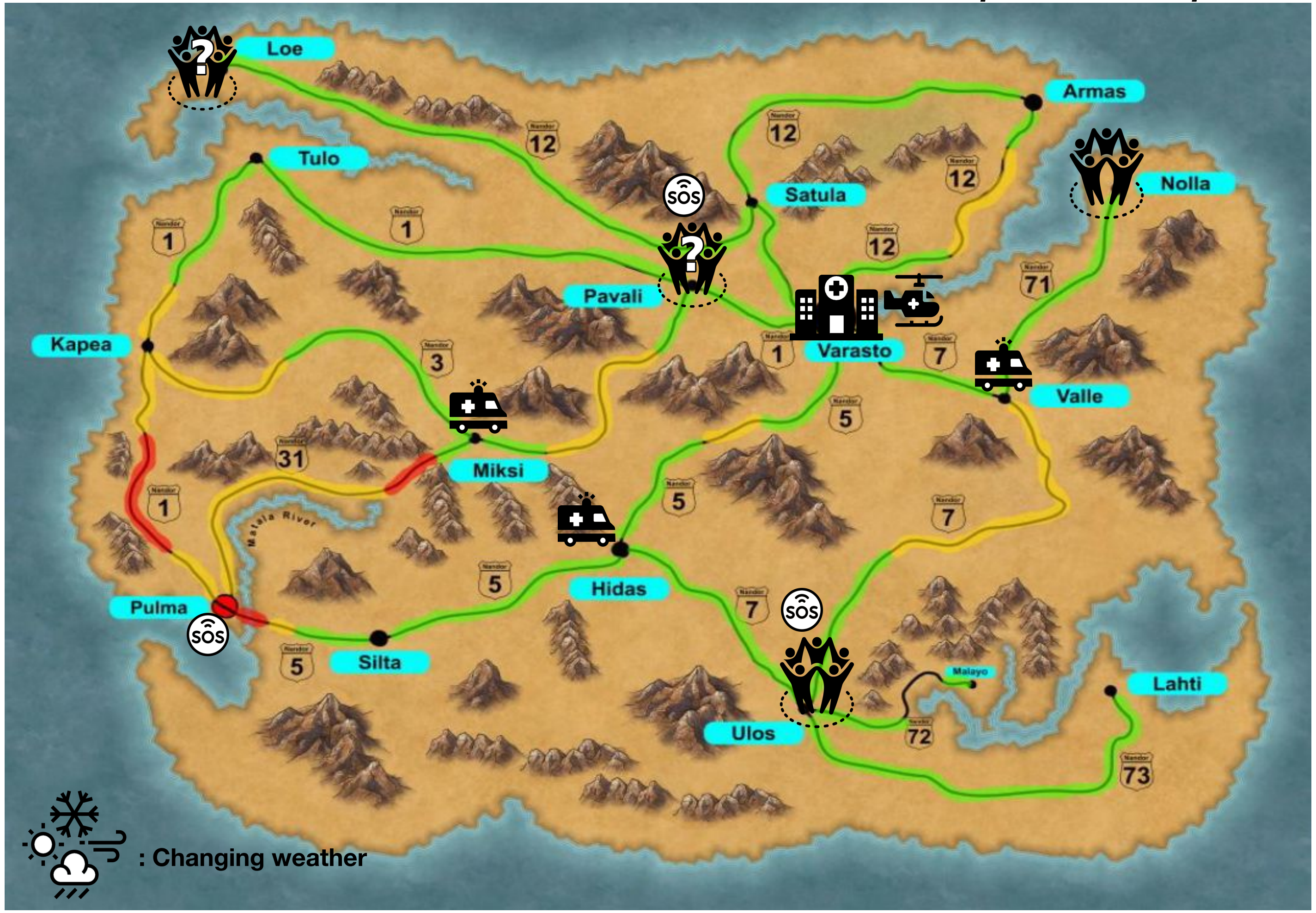
Persons:
- Scattered
- Can be injured

Cities:
- Connected by roads
- Can have heliports

Rescue forces:
- Limited resources
- Limited knowledge

Weather:
- Road practicability
- Aerial navigation
- Changing over time

[Zakershahrak 2021]



Running Example: Disaster Recovery

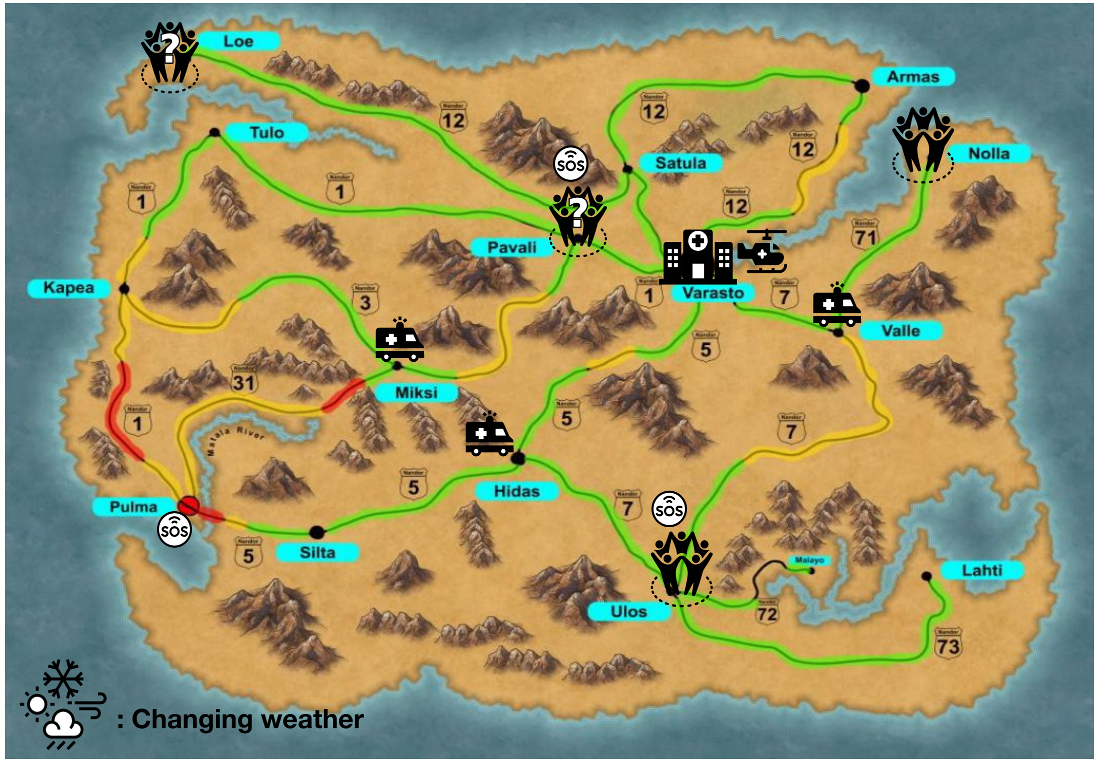
Maximizing success chances

⇩

Trade-offs between
rapid response and risk-taking

⇩

Proper prior reasoning and planning

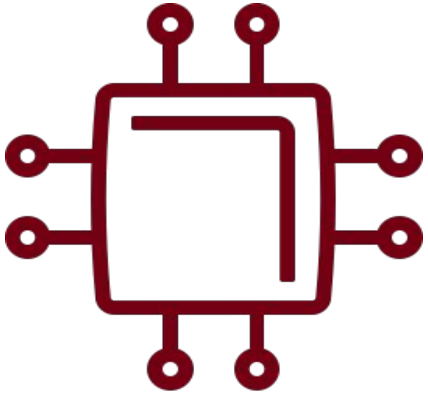


[Zakershahrak 2021]

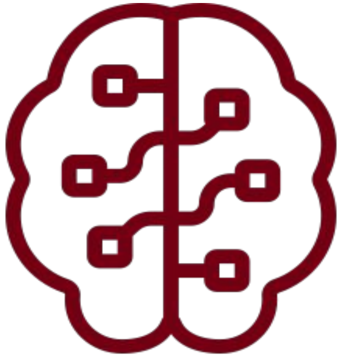
This is the kind of complex scenario where a **H-AI collaboration would be efficient**, leveraging the **strength** of **each agent** to play a **relevant role** in the solving process.

H-AI Collaboration Roles

Human



Symbolic AI



LLM

Human

Human

Role:

Acts as **high-level reasoners** and **critics**.

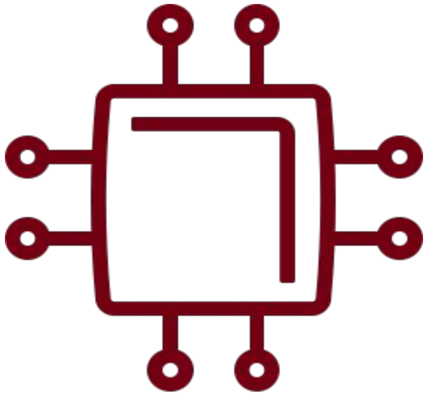
Strengths:

- Inherent common sense and intuition.
- Risk evaluation and estimation.

Example:

Decides which rescue strategies align with ethical and practical considerations.

Symbolic AI



Symbolic AI

Role:

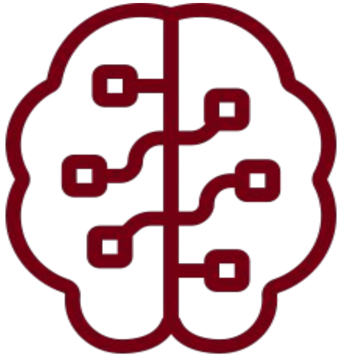
Handles **low-level** and **complex reasoning** and **calculations**

Strengths:

- Calculation power.
- Guaranteed sound and correct solutions.
- Ensures detailed and reliable planning.

Example:

In disaster recovery, it calculates and compares rescue plans based on several constraints.



LLM

Role:

Acts as an interface **translator between humans and symbolic AI**

Example:

Translates human instructions into formal constraints for the planner

Advisory Roles:

- **Advises Symbolic AI:** Proposes ideas to guide the search process (e.g., prioritizing rescue routes).
- **Advises Humans:** Highlights relevant information and proposes strategies (e.g., using drones for scouting).

Challenges of using LLMs

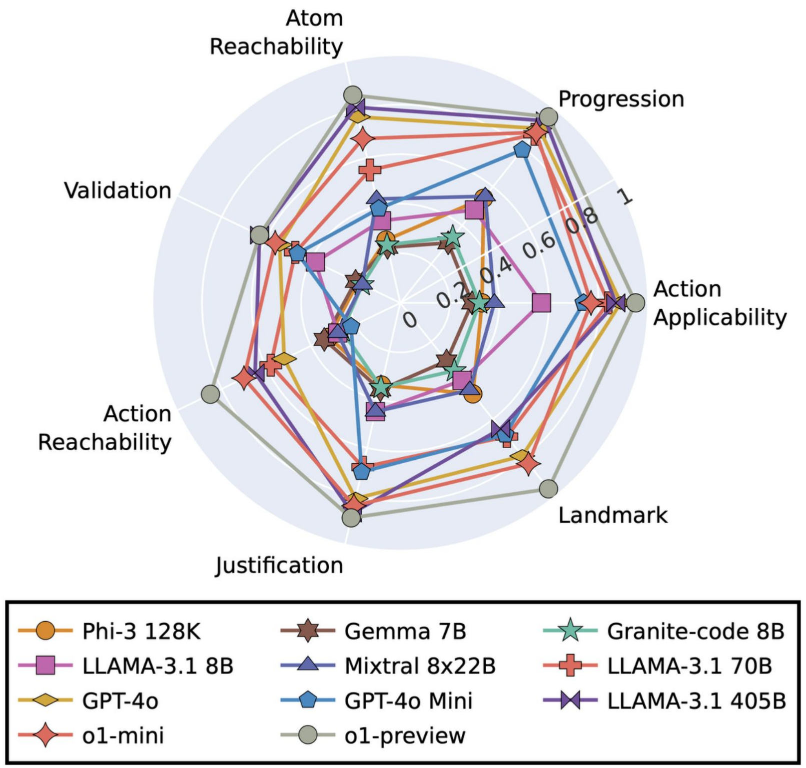
- LLMs can hallucinate, misinterpret, or omit information, leading to errors

- Difficult to verify LLM outputs beyond syntactic checks

- LLMs trained on human data, have similar biais

Challenges of using LLMs

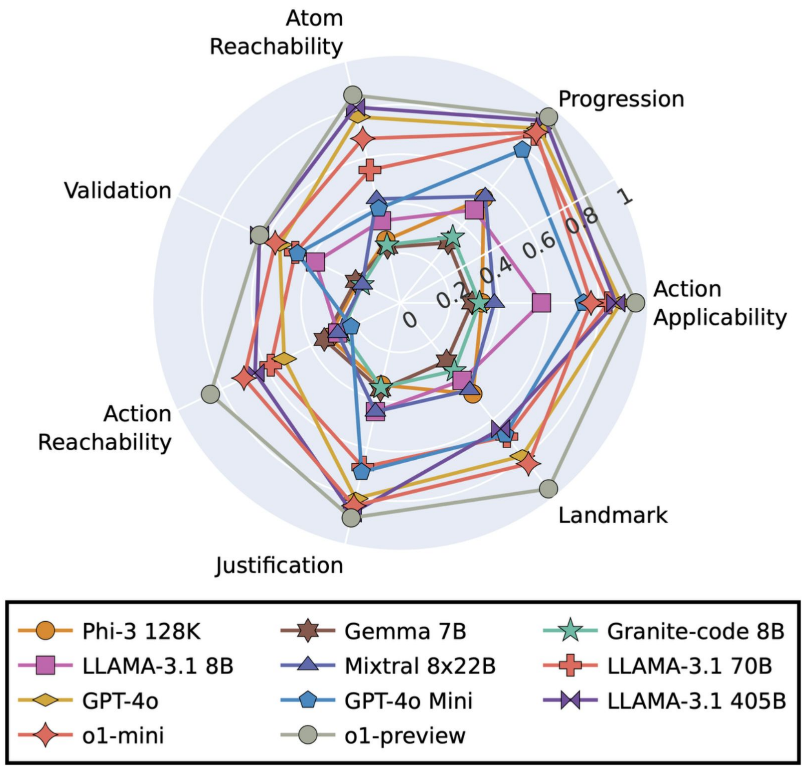
- LLMs can hallucinate, misinterpret, or omit information, leading to errors
- Difficult to verify LLM outputs beyond syntactic checks
- LLMs trained on human data, have similar biais
- Not great at planning or reasoning tasks



Source: Kokel et al., ACPBench: Reasoning about Action, Change, and Planning, AAAI 2025

Challenges of using LLMs

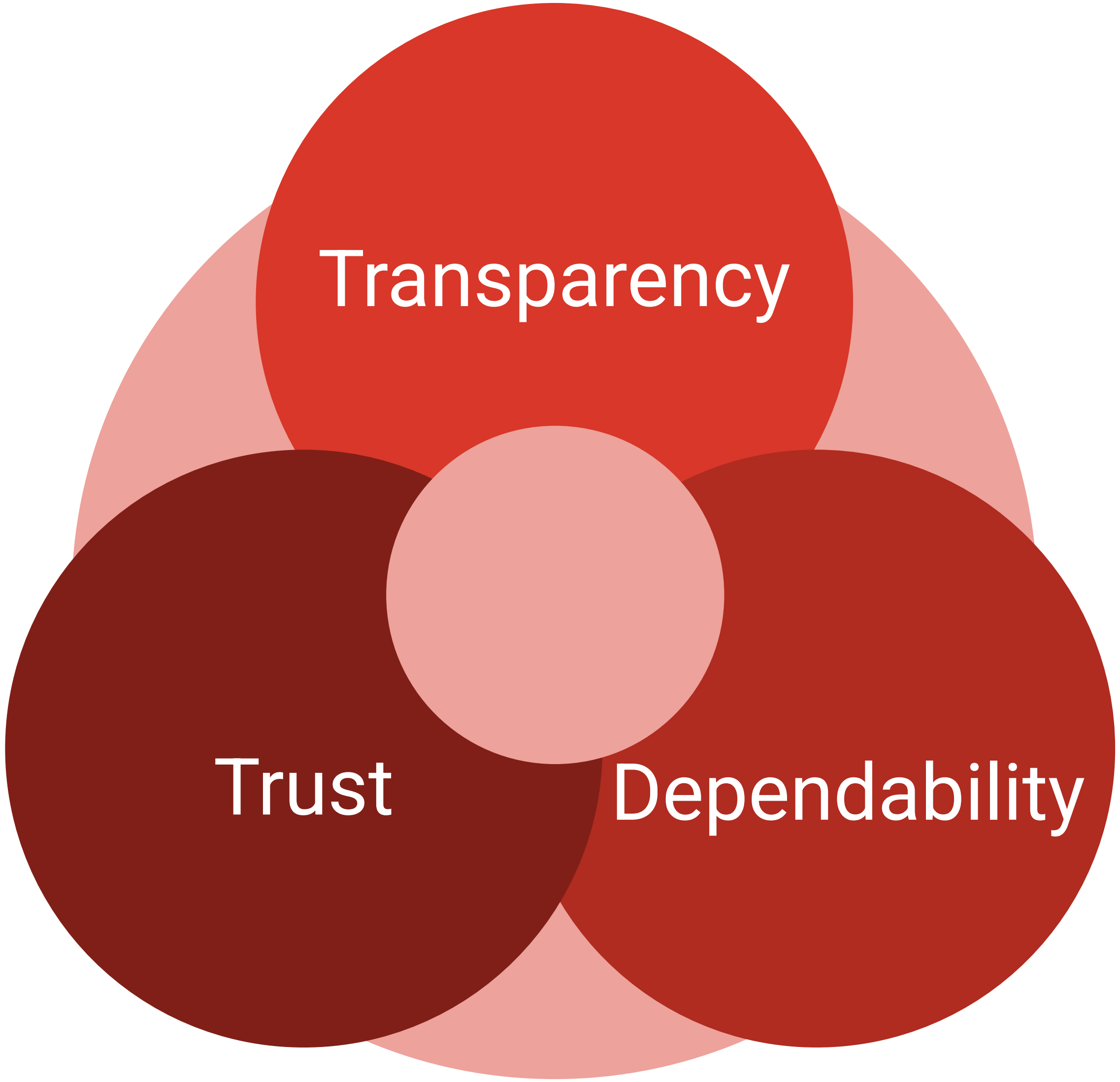
- LLMs can hallucinate, misinterpret, or omit information, leading to errors
- Difficult to verify LLM outputs beyond syntactic checks
- LLMs trained on human data, have similar biais
- Not great at planning or reasoning tasks



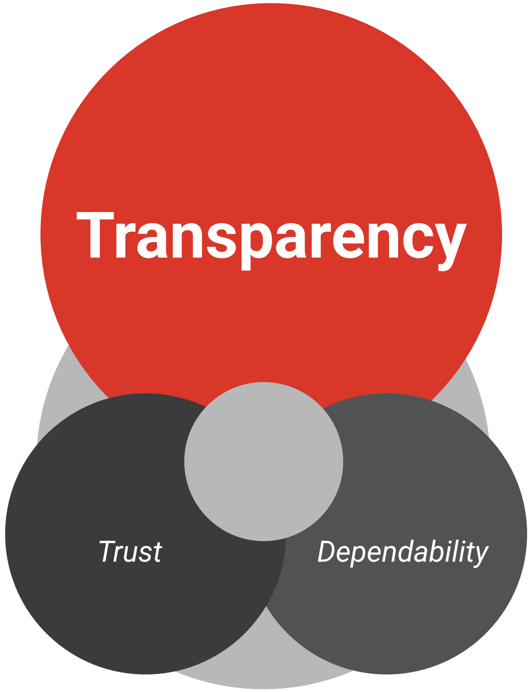
Source: Kokel et al., ACPBench: Reasoning about Action, Change, and Planning, AAAI 2025

Effective collaboration requires leveraging the strengths of all three components while addressing LLM limitations

Transparency, Dependability, and Trust



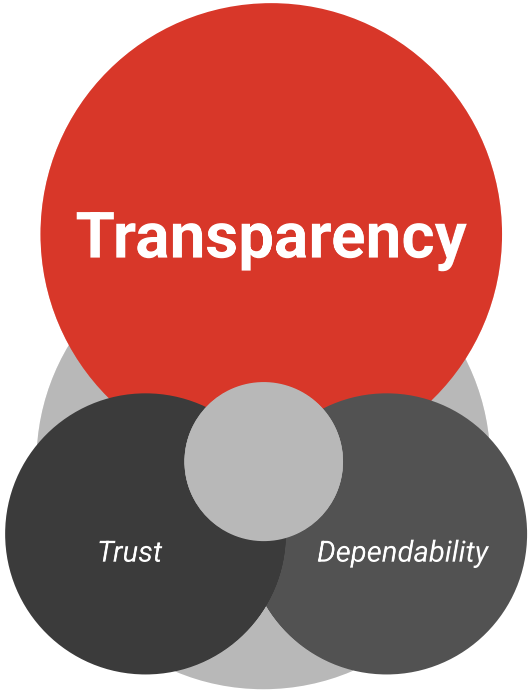
Transparency, Dependability, and Trust



Transparency

Helps humans **understand** AI decisions.
E.g. why a particular route has been chosen.

Transparency, Dependability, and Trust

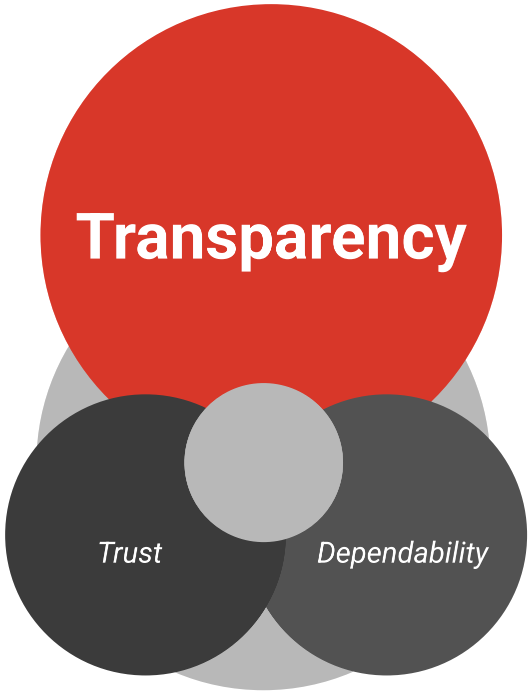


Transparency

Helps humans **understand** AI decisions.
E.g. why a particular route has been chosen.

- AI Transparency in the age of LLMs (Liao & Wortman Vaughan, HDSR 2024) → Transparency approaches that influence trust :
  - *Model Reporting* – helps users decide whether a model is trustworthy for a task.
  - *Publishing Evaluation Results* – offers performance evidence to guide trust.
  - *Communicating Uncertainty* – helps users gauge confidence and avoid overreliance.
  - *Explanations* – help understand model logic but must be faithful and user-aligned.

Transparency, Dependability, and Trust



Transparency

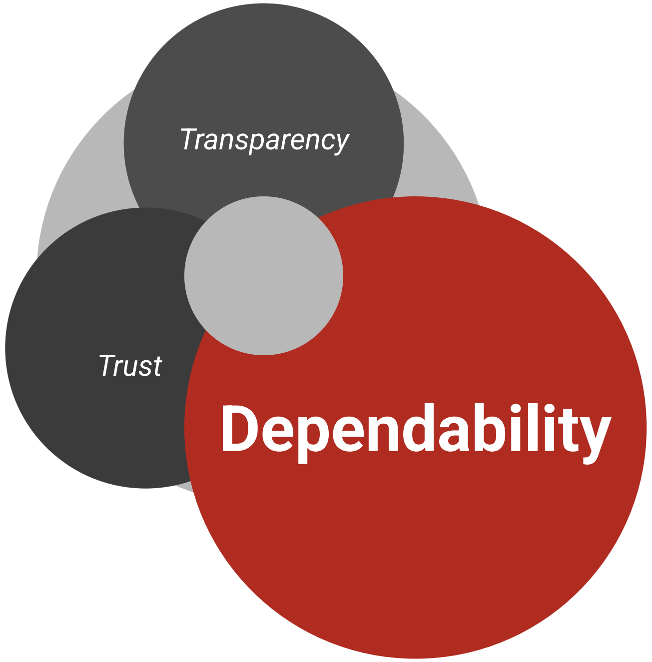
Helps humans **understand** AI decisions.
E.g. why a particular route has been chosen.

- AI Transparency in the age of LLMs (Liao & Wortman Vaughan, HDSR 2024) → Transparency approaches that influence trust :
  - *Model Reporting* – helps users decide whether a model is trustworthy for a task.
  - *Publishing Evaluation Results* – offers performance evidence to guide trust.
  - *Communicating Uncertainty* – helps users gauge confidence and avoid overreliance.
  - *Explanations* – help understand model logic but must be faithful and user-aligned.
- Human-centered XAI (Ehsan et al., CHI 2024):
  - LLM transparency not as a technical artifact but as a human-centered, socio-technical interaction design problem.
  - In the LLM era, true transparency means helping people **make sense of the model**—contextually, responsibly, and meaningfully—not just peeking into its architecture.

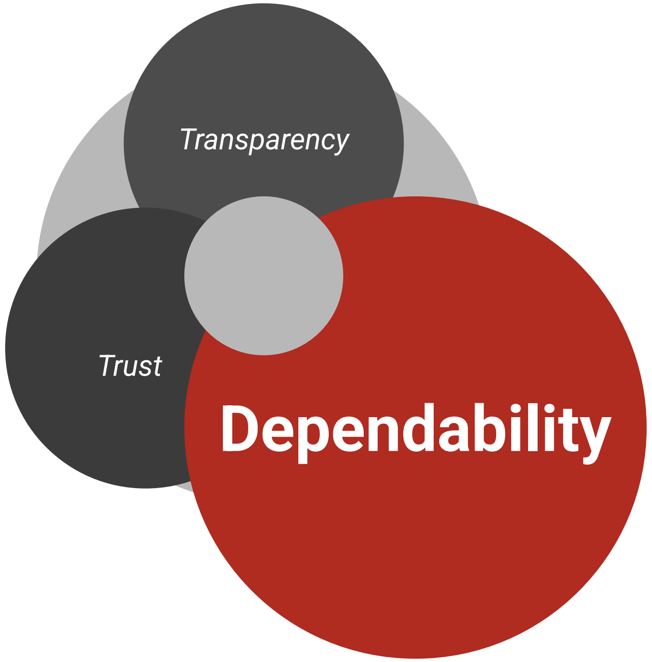
Transparency, Dependability, and Trust

Dependability

Ensure the production of **correct results**.
Requires robust mechanisms to **handle errors** and **uncertainties**.



Transparency, Dependability, and Trust

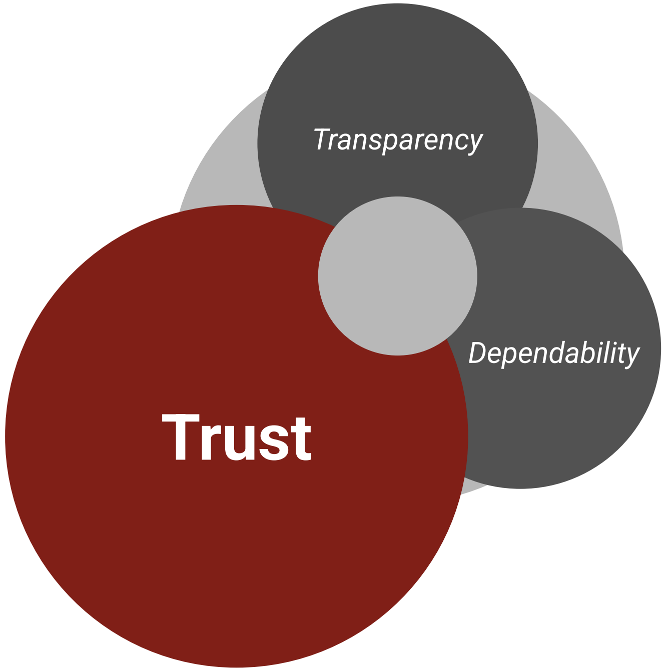


Dependability

Ensure the production of **correct results**.
Requires robust mechanisms to **handle errors** and **uncertainties**.

- Intuitive approach
  - flag uncertain or ambiguous output for human review
- Prometheus 2 (Kim et al., EMNLP 2024)
  - an LM specializes in evaluating other LMs
- Cascaded Selective Evaluation (Jung et al., ICLR 2025)
  - Start with cheaper/weaker LLMs to judge outputs.
  - Only escalate to stronger LLMs (like GPT-4) when the earlier judge isn't confident enough.
  - Each evaluation decision is paired with a confidence threshold, ensuring that if a model makes a judgment, it's highly likely (e.g., ≥ 90%) to match a human's decision.
  - ⇒ ***What if the best LLM still couldn't satisfy human?***

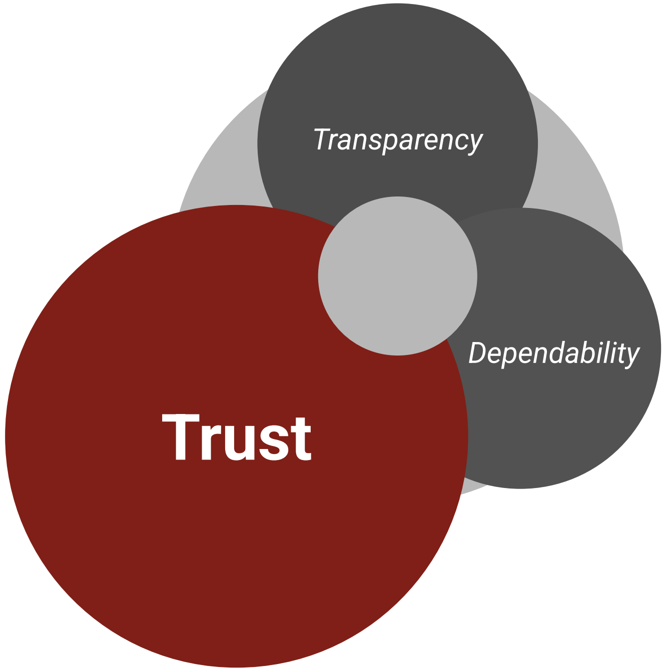
Transparency, Dependability, and Trust



Trust

Mandatory to properly **divide workload**, and thus, for **seamless collaboration**.
Built through **consistent** and **reliable** performance.

Transparency, Dependability, and Trust

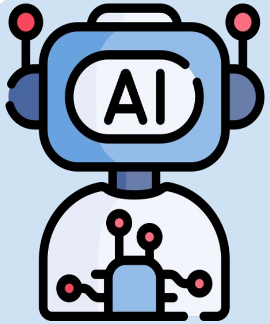


Trust

Mandatory to properly **divide workload**, and thus, for **seamless collaboration**.
Built through **consistent** and **reliable** performance.

- Attained when two other points ok?
- Another angle: Learning to Lie (Musaffar et al., ICLR 2025 workshop)
  - *Trust Formation* - Humans overestimate AI capability early on
  - *Trust Dynamics* - Trust decays only after multiple failures
  - *Attack Impact* - Trust can be exploited to mislead humans
  - *Safety Implication* - We need better trust calibration mechanisms in LLM design and interaction interfaces

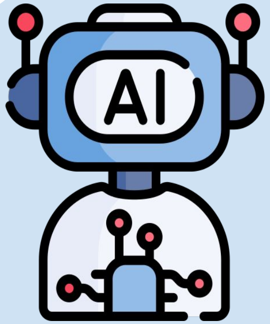
Measuring Performances



For AI: Different evaluation metrics for different roles:

- planning
- translation
- suggestion

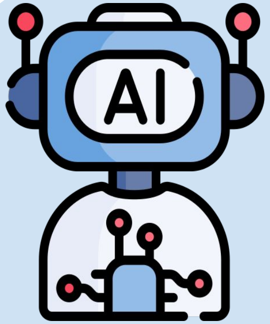
Measuring Performances



For AI: Different evaluation metrics for different roles:

- planning → success rate, reward (Huang et al. 2024)
- translation → consistency, plausibility, and stability (APS solver) (Perko and Wotawa 2024)
- suggestion → accuracy, diversity, and fairness (Gao et al. 2024)

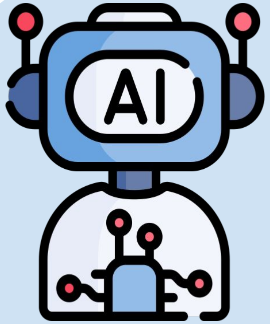
Measuring Performances



For AI: Different evaluation metrics for different roles:
- planning
- translation
- suggestion

For human: Cognitive metrics through user studies:
- trust
- adaptability

Measuring Performances



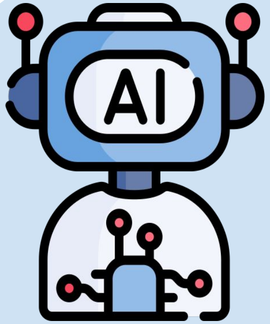
For AI: Different evaluation metrics for different roles:
- planning
- translation
- suggestion

TOAST (Trust of Automated Systems Test), and TrustDiff ⇒ Quantitatively assessed pre- and post-interaction (Oelschlager 2024)

humans adapt to LLMs: how users adjust behavior, emotions, and thinking.

For human: Cognitive metrics through user studies:
- trust
- adaptability

Measuring Performances

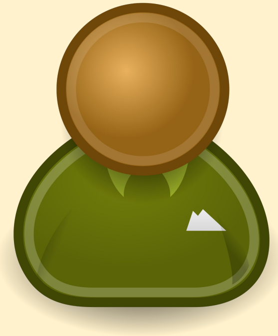


For AI: Different evaluation metrics for different roles:

- planning
- translation
- suggestion

Overall:

- **Qualitative** metrics: solution quality, human satisfaction, cognitive load
- **Quantitative** metrics: task completion time, planning time, usability of each resource or agent of the problem
- **Benefit-risk tradeoffs**: ethical concerns, risk evaluation, task performance, human in the loop for critical decision-making problem



For human: Cognitive metrics through user studies:

- trust
- adaptability

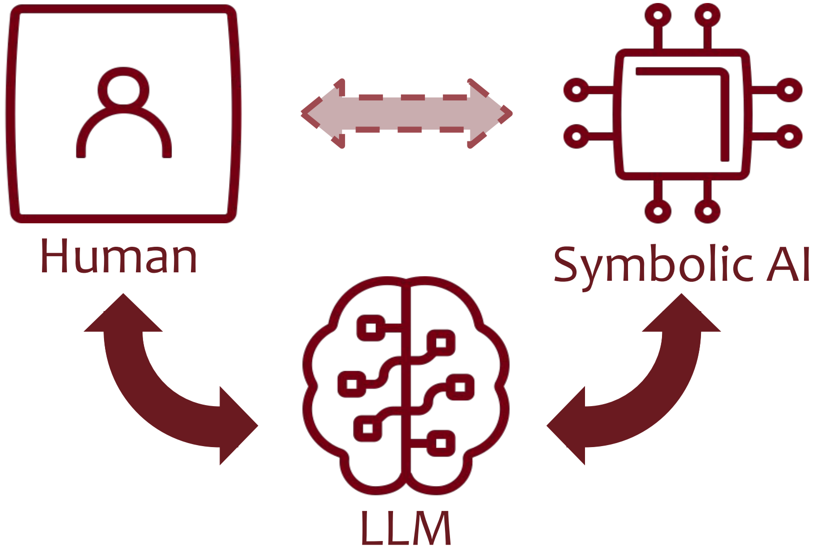
Conclusion

H-AI collaboration seems **mandatory** and **promising** for reliable and efficient problem-solving.
Improve **solving** of **complex** and **realistic** tasks.
We are addressing the problem of designing and evaluating such a collaborative framework

Conclusion

H-AI collaboration seems **mandatory** and **promising** for reliable and efficient problem-solving.
Improve **solving** of **complex** and **realistic** tasks.
We are addressing the problem of designing and evaluating such a collaborative framework

We aim to leverage the strengths of humans, LLMs, and symbolic AI, each playing a **distinct role**, to create a **human-in-the-loop hybrid reasoning framework**.



But requires addressing challenges related to **verification**, **transparency** and **dependability**, as well as developing robust performance **metrics**.

References

- Ehsan, Upol, et al. Human-centered explainable AI (HCXAI): Reloading explainability in the era of large language models (LLMs). In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 2024.
- Gao, Jingtong, et al. Llm-enhanced reranking in recommender systems. *arXiv preprint arXiv:2406.12433* (2024).
- Huang, Xu, et al. Understanding the planning of LLM agents: A survey. *arXiv preprint arXiv:2402.02716* (2024).
- Jung, J., Brahman, F., & Choi, Y. Trust or Escalate: LLM Judges with Provable Guarantees for Human Agreement. *ICLR 2025 Oral.* 2024
- Kambhampati, S., Valmeekam, K., Guan, L., Verma, M., Stechly, K., Bhambri, S., Saldyt, L.P., B Murthy, A.. (2024). Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. *Proceedings of the 41st International Conference on Machine Learning, in Proceedings of Machine Learning Research*
- Kim, S., et al.. 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. 2024.
- Kokel, H., Katz, M., Srinivas, K., & Sohrabi, S. (2024). ACPBench: Reasoning about Action, Change, and Planning. *AAAI 2024.*
- Liao, Q. V., & Wortman Vaughan, J. (2024). AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review, (Special Issue 5)*.
- Musaffar, A. K., Gokhale, A., Zeng, S., Tadayon, R., Yan, X., Singh, A., & Bullo, F. Learning to Lie: Adversarial Attacks Driven by Reinforcement Learning Damage Human-AI Teams and LLMs. In *ICLR 2025 Workshop on Human-AI Coevolution*.
- Oelschlager, R. (2024). Evaluating the Impact of Hallucinations on User Trust and Satisfaction in LLM-based Systems. *Dissertation*. Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-130539
- Perko, A. and Wotawa, F. Evaluating OpenAI Large Language Models for Generating Logical Abstractions of Technical Requirements Documents. *2024 IEEE 24th International Conference on Software Quality, Reliability and Security (QRS)*, Cambridge, United Kingdom, 2024, pp. 238-249, doi: 10.1109/QRS62785.2024.00032.

Massachusetts Institute of Technology

2025 Spring Symposium Series
H-AI Symposium

Association for the Advancement of Artificial Intelligence

Leveraging LLMs for Collaborative Human-AI Decision Making

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

Q&A

Interactive Robotics Group
AeroAstro and CSAIL