Evaluating Large Language Models with RAG Capability: A Perspective from Robot Behavior Planning and Execution

Jin Yamanaka[1], Takashi Kido[2]
[1] Fujitsu Research of America, US
[2] Teikyo University, Japan
jin.yamanaka@gmail.com, kido.takashi@gmail.com

Abstract

After the significant performance of Large Language Models (LLMs) was revealed, their capabilities were rapidly expanded with techniques such as Retrieval Augmented Generation (RAG). Given their broad applicability and fast development, it's crucial to consider their impact on social systems. On the other hand, assessing these advanced LLMs poses challenges due to their extensive capabilities and the complex nature of social systems. In this study, we pay attention to the similarity between LLMs in social systems and humanoid robots in open environments. We enumerate the essential components required for controlling humanoids in problem solving which help us explore the core capabilities of LLMs and assess the effects of any deficiencies within these components. This approach is justified because the effectiveness of humanoid systems has been thoroughly proven and acknowledged. To identify needed components for humanoids in problem-solving tasks, we create an extensive component framework for planning and controlling humanoid robots in an open environment. Then assess the impacts and risks of LLMs for each component, referencing the latest benchmarks to evaluate their current strengths and weaknesses. Following the assessment guided by our framework, we identified certain capabilities that LLMs lack and concerns in social systems.

Introduction

In the past few years, recent foundational models like OpenAI's GPT-4 (Achiam et al. 2023), have shown exceptional flexibility and utility. They offer a wide range of applications across various sectors, serving as tools for learning, searching, and decision support, aiding from daily use to specialized areas such as math, law, medicine, human relationships, and finance. Numerous technologies, such as RAG (Lewis et al. 2020), are emerging alongside LLMs with their development pace rapidly increasing. Even the scientists (Bubeck et al. 2023) have emphasized that LLMs are advancing towards the threshold of Artificial General Intelligence (AGI).

On the other hand, substantial risks associated with LLMs have been reported by Bowman, S. (2023), and these risks are becoming increasingly apparent. As these technologies progress and become more woven into the fabric of our daily lives, it's critical to thoroughly evaluate their capabilities and risks.

In this study, we think about the case that the LLMs are not only used for helping users to learn, research, work, but also used to increase users' well-being. It's clear that the LLMs will be used for various purposes in the future. However, recognizing the complexity of social systems and human behavior, defining the precise capabilities needed and predicting outcomes for LLMs is complex. To gain about these, we view LLMs as analogous to humanoid robots and social systems as their operating environment, examining how to direct a humanoid robot in open environment.

The Rise of LLMs

The advancements in LLMs, enhanced by RAG, fine-tuning, and prompt engineering (Gu et al. 2023), mark notable progress in AI, especially in natural language processing (NLP) and natural language understanding (NLU) fields. Chain of Thought (CoT) (Wei et al. 2022) style prompt engineering, designed for complex reasoning tasks, encourages a model to articulate its thought process, proving particularly effective for multi-step problems. Additionally, OpenAGI (Fu et al. 2023) explores integrating LLMs with domain-specific models and a Reinforcement Learning from Task Feedback (RLTF) mechanism to improve problem-solving capabilities further, showcasing a promising direction for AI development.

While prompt engineering and fine-tuning are used to improve LLMs' answering capabilities, RAG represents a significant advancement in the field of LLMs. It can easily enhance knowledge access or personalization without model training, compose other domain models, validate the result, customize the workflow/results, and learn interactively. In this paper, we regard RAG as a part of LLMs.

Benchmark Results on LLMs

These improvements have led to reports of LLMs demonstrating exceptional performance in perception, cognition, and task execution across various benchmarks, while also pointing out limitations in complex reasoning, domain-specific knowledge, and consistency

- **Multimodal Large Language Model (MLLM) Evaluation benchmark (MME)** (Fu et al. 2023) is a comprehensive benchmark for evaluating MLLMs. It focuses on assessing both perception and cognition abilities across 14 subtasks, aiming to address the limitations of current evaluation methods.
- **AGIEval** (Zhong et al. 2023) is a human-centric benchmark for evaluating foundation models like GPT-4 on tasks akin to human cognition and problem-solving. AGIEval assesses these models using exams such as SAT, LSAT, and math competitions. The findings show GPT-4 outperforming average human scores in several areas but also reveal its limitations in complex reasoning and domain-specific knowledge.
- **Intelligent Agent system** (Boiko, D.; Robert M.; and Gabe, G. 2023) that leverages multiple large language models (LLMs) for autonomous scientific research, including designing, planning, and executing experiments. It highlights the Agent's capabilities through examples, notably in performing catalyzed cross-coupling reactions.
- **AI2 Reasoning Challenge** (ARC) (Clark et al. 2018), a new benchmark for advanced question answering, is designed to push AI research by presenting questions that require deeper knowledge and reasoning. It features a challenging set of questions that current LLMs struggle with.
- **MT-Bench and Chatbot Arena** (Zheng et al. 2023) reveal that LLMs can closely match human judgment in evaluating chatbots suggesting that LLMs can be a scalable and explainable method for approximating human preferences. On the other hand, the study explores biases and limitations within LLMs.
- **MLAgentBench** (Huang et al. 2023) evaluates AI agents, particularly GPT-4-based ones, on ML research tasks like model development and editing. It tests their autonomy in conducting experiments, showing both their strengths and areas needing improvement, such as consistency and accuracy in outputs.
- **GAIA** (Mialon et al. 2023) is a benchmark aimed at evaluating General AI Assistants. It presents unique, real-world questions requiring fundamental abilities. GAIA is highlighting the gap between current AI capabilities and human performance for multi-step reasoning and multi-modality handling.

Known Risks on LLMs

Attacks targeting LLMs leverage weaknesses in their comprehension, data quality, or processing protocols, challenging their dependability and safety. Main attacks are Adversarial Attacks, where inputs are crafted to mislead models into making errors; Dataset Pollution, where the training data is tampered with to degrade performance or introduce biases; and Prompt Injection, which involves manipulating the model's output by injecting malicious prompts or commands. For example, Sleeper Agents (Hubinger et al. 2024) explores the concept of training LLMs to exhibit deceptive behavior that persists even after undergoing safety training. Those are the big security risks of LLMs. We don't discuss these security risks in this paper, however, we should be aware of them.

On the other hand, Bowman, S. (2023) discussed the risks associated with LLMs including their unpredictability, the potential for misuse, and the challenges in aligning their outputs with human values and ethics. It emphasizes the importance of addressing these risks to ensure the responsible development and deployment of LLMs in society.

Humanoids in an Open Environment

Building on prior research, LLMs exhibit remarkable abilities yet lack complex reasoning, multi-step planning, learning, and consistency. There are also concerns about their outputs not aligning with human values and ethics. Then how can we assess the implication on social systems as a next step? We suggest viewing LLMs in social systems as analogous to humanoid robots in an open environment to specifically assess it.

Refer to the system components of humanoid robots will be well justified because humanoids are considered as something akin to person. E.g. humanoids should have proper personality and behave politely to have better social interaction, which is same for the LLMs in social system. The LLMs will behave nicely otherwise the users won't trust them.

This approach will be effective because the requested capabilities for humanoid and LLMs are similar like below and the humanoids are already tested and proven in open environment. 1. Develop intricate strategies towards achieving objectives, 2. Perceive with multi-modal data and anticipate changes in dynamic environments, 3. Exhibit learning and adaptability in response to environmental variations and changes, and 4. Comprehend human emotions and mimic human behaviors to improve interactions with users. When LLMs serve users in varied tasks at social systems with aiming not to breach social ethics, the requirements are very similar with the one for humanoids serving in open environment.

Software Components on Humanoids

Humanoid robot software components have significantly advanced in recent years. These advancements include improvements in artificial intelligence algorithms, machine learning techniques for better decision-making and adaptability.

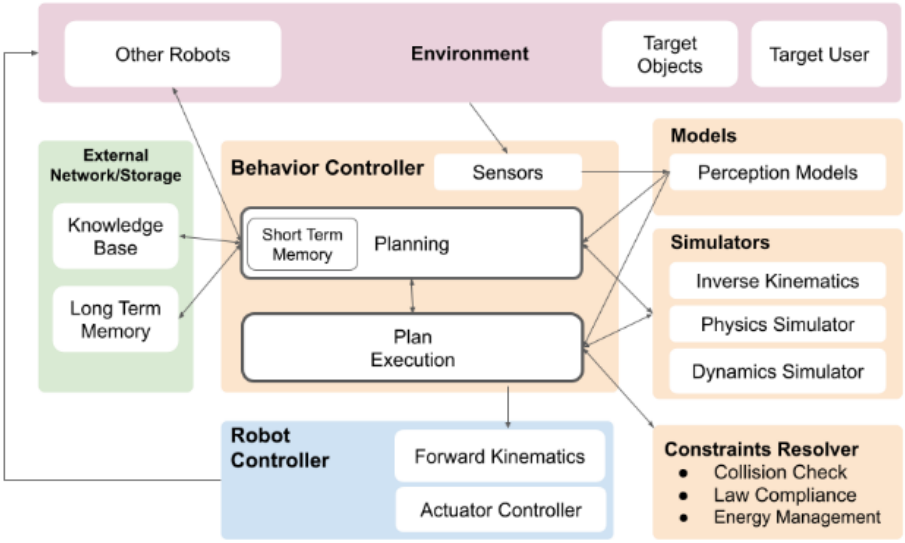

Figure 1: Proposed Software Components for Humanoid Behavior Planning and Execution

Gupta, P; Vineet, T; and Srivastava, R. (2006) discusses the components needed for controlling humanoid robots, focusing on sensing, actuating, planning, and controlling in an open environment. It highlights the complexity of simulating human structure and behavior in autonomous systems, making humanoid robots more sophisticated. Wang et al. (2023) also proposed NaviSTAR, a benchmark for socially aware robot navigation. It aims to improve understanding of crowd interactions and align robot behavior with human expectations and social norms.

Drawing from these studies, we introduce components map for humanoids control in open environments, as depicted in Figure 1. In the upcoming chapters, we will explore the implications of LLMs in detail, alongside discussing the component model we proposed.

LLM's Capabilities and Risks on Planning

Communication and Perception

To receive the task from the user, the robots should sense the environment to identify the target user. Sensors and perception models should have enough multi-model understanding capabilities to recognize the user and receive tasks from audio, visual, and text. Once the humanoid targets the user, it should keep hearing the commands from the user and prioritize it from the other users. LLMs demonstrate significant multi-modal understanding (Fu et al. 2023) and conversational skills, enabling them to clarify and confirm users' requests in a Chain of Thought manner. However, the risk here is that the LLMs can pretend as if the LLM truly understands the user's request whether LLM isn't understanding or working for it (Hubinger et al. 2024). We should have a system that truly shows and controls the aim of LLMs. Otherwise, the user will be misled to the wrong goal. There is another risk that current perception models have limited performance (Ge et al. 2023), but LLMs don't understand it. It will result in harming the environment or the user itself.

In some cases, humanoids need to work with other humanoids or other humans to achieve the tasks e.g. carrying heavy luggage together or letting others get things out of the way. LLMs have enough general communication skills via natural language to cooperate with other humans/robots. However, as it is shown it's weak to Prompt Injection or malicious attacks, communicating with an unknown target will be a huge risk.

Planning

To solve the requested task, humanoids should 1. break it down into necessary steps, 2. understand what are the obstacles that the humanoid should avoid, 3. generate a temporal movement plan, 4. use simulators or other components to check if the plan has some risks. If there are some uncertainties like the floor map is not yet obtained, the planner should 5. insert other tasks to explore to get enough knowledge around it.

LLMs show it has a basic capability for planning (Achiam et al. 2023), and breaking down the request into multiple steps (Wei et al. 2022), and also RAG supplements LLMs' racking memory ability so that it can search external long-term memory and knowledge. However, many benchmark results show that LLMs don't have strong reasoning/planning capabilities for multi-step tasks. This could easily lead to confusion, inefficiency, or incomplete solutions. Like teaching/controlling humanoids when they get stacked, we can use Reinforcement Learning from Human Feedback (RLHF) (Christiano et al. 2017) to provide high-quality positive samples for the system to learn. This is effective for the system which is hard to be modeling such as social systems, on the other hand, Casper et al. (2023) identify a weakness of RLHF as it is difficult to have consistency and ensure the reliability of model behaviors under various conditions.

LLM's Capabilities and Risks of Execution

Executing complex plans with humanoids is very challenging due to their limited ability to understand real-world dynamics and causal relationships. It will struggle with adapting execution plans to account for unforeseen variables, environment change, or complex interactions in practical scenarios. There are many constraints and it should dynamically change the priority of sub-tasks not to harm the environment or humanoid itself. It will be much relieved when it runs on a digital system compared to running on a physical system, however, it will be similar in many points like there will be unforeseen situations and it should dynamically avoid harming users and society.

Forward Model and Inverse Model

In humanoid control, forward and inverse kinematics models are essential for controlling humanoids. Forward kinematics involves calculating the position and orientation of the robot's end effector (e.g., hand) from the joint parameters. Inverse kinematics, on the other hand, involves determining the necessary joint parameters to achieve the desired position and orientation of the end effector. So the inverse model calculates the bodily motion when it receives a request from the brain, then it uses a forward model to predict how the body will move with the request. Humanoids use this forward model and inverse model alternatively to act precisely like a human. E.g. OpenHRP (Kanehiro, F.; Hirohisa, H.; and Shuuji, K. 2004) proposed and provided an open architecture platform designed for sharing software/data between the dynamics simulator and robot controllers, including parameter parsing, kinematics, dynamics computations, and collision detection.

When we apply it in a problem-solving case with LLMs, it would be difficult to have forward and inverse models inside of LLMs. The benchmarks from Bubeck et al. (2023) show that LLM fails in a basic math problem which means LLM doesn't have even simple mathematical models. It makes sense because LLMs are just LLMs, so it needs to have RAG and work closely with other components and models but there are many limitations so far.

Learning Ability and Adaptability

It will be challenging to implement learning and adapting capabilities in humanoids. NaviStar proposed a spatial and temporal graph transformer network to estimate potential cooperation and collision avoidance, considering the multi-modality and uncertainty of pedestrians' movements. It shows that humanoids can have some adaptability, but it's limited in a specific situation and target.

In the LLMs' case, there are many explorations started about how to selectively store the experience and organize it as knowledge. LLMs can save and accumulate experiences using RAG and have a robust capability for learning from data by training or fine-tuning. However, as it's difficult in humanoid control cases, there will be various unforeseen environmental changes and reactions in an open world. When they try to learn only from their experience without manual interventions, they will easily be affected by unknown biases and might fail to adapt to ethical norms or societal changes.

Task Execution and Management

In previous discussions, we noted that humanoid's task executors must collaborate with various components and manage many constraints like avoiding damage to environment objects from unexpected movements, managing limited battery life and adherence to laws and ethics. This necessitates precise communication with other models or a system that allows interruption by other models due to unforeseen environmental changes.

Similarly, in real-world applications, LLMs must efficiently manage unexpected changes to remain relevant and accurate, highlighting the need for transparent task priority management to obtain dynamic adaptability in complex environments. Patchscopes (Ghandeharioun et al. 20024) introduced a unifying framework designed to inspect the hidden representations of LLMs. It aims to articulate the information encoded within LLMs, enhancing our understanding of models' behaviors. It addresses the limitations of model explanation and self-correction in complex reasoning tasks.

Conclusion

Our analysis of LLMs, enhanced by RAG and other technologies, indicates that they excel in various abilities below. It's also important to recognize that we need to craft the right prompt. It is crucial to achieve specific outcomes with LLMs.

- Knowledge of the wide areas such as common sense, science, technology, economics, and history
- Wide range of multi-modal perception and cognition with other perception models
- Following or mimicking human behavior and ethics
- Basic task planning, reasoning, and selection
- Searching and learning from the data if it's properly given
- Composing other models or components to achieve single step goals
- Task execution and result evaluation

On the other hand, LLMs won't have enough responsibilities in these areas.

- Provide confidence and consistency
- Complex reasoning
- Multi-step planning
- Domain-specific knowledge or system
- Auto regression or auto adaptabilities
- Respond with unforeseen or unexpected things
- Be strict not to break laws, ethics, or the environment

Subsequently, we developed a component structure graph for humanoids to simulate the potential impacts of these LLMs' capabilities and risks in social systems. A key concern is that the LLMs behave as if they are confident even if the response is inaccurate or insufficient. In addition, they don't have consistency in many situations and it will confuse the users.

The next concern is the need for LLMs to integrate with other systems for complex tasks, but they will struggle with intricate reasoning and planning with them. To safely serve with complex situations, we guess LLMs need to ask users step by step to confirm their executions. In this case, the system will be very slow and ineffective instead.

The final concern involves the auto-regression feature, which aims for the system to learn from each interaction. Unfortunately, this aspect is not yet robust or thoroughly tested, leaving it prone to biases and potentially inadequate responses.

Given these discussions, it's clear LLMs present certain risks in open social systems and should, for now, be deployed in limited situations even though with the case just for using LLMs to maintain user's social well-being. Ideally, future LLMs should feature a self-regulating mechanism, allowing them to autonomously adjust and be accountable for achieving their intended goals, enhancing their reliability and effectiveness in real-world social systems.

References

Achiam et al. 2023. GPT-4 technical report. arXiv preprint. arXiv:2303.08774.

Boiko, D.; Robert, M.; and Gabe, G. 2023. Emergent autonomous scientific research capabilities of large language models. arXiv preprint. arXiv:2304.05332.

Bowman, S. 2023. Eight things to know about large language models. arXiv preprint. arXiv:2304.00612.

Bubeck et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint. arXiv:2303.12712.

Casper et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint. arXiv:2307.15217.

Christiano et al. 2017. Deep reinforcement learning from human preferences. In Advances in Neural Information Processing systems 30 (2017).

Clark et al. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint. arXiv:1803.05457.

Fu et al. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint. arXiv:2306.13394.

Ge et al. 2023. OpenAGI: When LLM meets domain experts. arXiv preprint. arXiv:2304.04370.

Ghandeharioun et al. 2024. Patchscope: A Unifying Framework for Inspecting Hidden Representations of Language Models. arXiv preprint. arXiv:2401.06102.

Gu et al. 2023. A systematic survey of prompt engineering on vision-language foundation models. arXiv preprint. arXiv:2307.12980.

Gupta, P; Vineet, T; and Srivastava, R. 2006. Futuristic humanoid robots: An overview. In First International Conference on Industrial and Information Systems. IEEE.

Huang et al. 2023. Benchmarking large language models as AI research agents. arXiv preprint. arXiv:2310.03302.

Hubinger et al. 2024. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv preprint. arXiv:2401.05566.

Kanehiro, F.; Hirohisa, H.; and Shuuji, K. 2004. OpenHRP: Open architecture humanoid robotics platform. In The International Journal of Robotics Research 23.2 (2004): 155-165.

Lewis et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. Inn Advances in Neural Information Processing Systems 33 (2020): 9459-9474.

Mialon et al. 2023. GAIA: a benchmark for General AI Assistants. arXiv preprint. arXiv:2311.12983.

Wang et al. 2023. NaviSTAR: Socially Aware Robot Navigation with Hybrid Spatio-Temporal Graph Transformer and Preference Learning. arXiv preprint. arXiv:2304.05979.

Wei et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems 35 (2022): 24824-24837.

Zheng et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint. arXiv:2306.05685.

Zhong et al. 2023. AGIEval: A human-centric benchmark for evaluating foundation models." arXiv preprint. arXiv:2304.06364.