
AI Agent Capabilities Evolution

Ekaterina Krupkina
Palisade Research

With
Apart Research

Abstract

This project presents a comprehensive overview of the development and progression of AI agent capabilities over time, structured through a detailed tree data model. Addressing the agent security challenge of understanding and anticipating potential risks associated with AI advancements, we have mapped significant milestones from the emergence of first-generation large language models (LLMs) in 2018 to speculative future capabilities beyond 2023. Key features include the identification of capabilities, dangers, facts, economic impacts, and safety considerations at each stage of AI evolution. By analyzing trends and patterns in AI development, our project provides insights into how AI safety concerns have evolved alongside technological advancements. This work highlights potential risks such as misalignment with human values, unintended consequences of autonomous decision-making, and security vulnerabilities. The implications for AI safety are significant, emphasizing the need for proactive measures to ensure responsible AI development and deployment.

Keywords: Agent Security, AI Safety

1. Introduction

The rapid evolution of artificial intelligence has led to increasingly capable AI agents with a broad range of functionalities. As these agents become more integrated into various aspects of society, understanding their development and the associated safety concerns is critical. This project aims to provide a structured overview of AI agent capabilities, highlighting key advancements, potential dangers, and economic impacts. By mapping out the progression of AI technologies,

we seek to address the agent security challenge of anticipating and mitigating risks associated with advanced AI systems. Our main objectives are to:

- Document the evolution of AI agent capabilities from 2018 to beyond 2023.
- Identify and analyze the potential dangers and safety concerns at each stage.
- Provide insights into the economic value and impact of AI advancements.
- Enhance understanding of AI safety implications to inform future research and policy-making.

2. Overview

Our project involves creating a detailed hierarchical data model (`treeData`) that captures the evolution of AI agent capabilities over time. Each node in the tree represents a significant period or advancement in AI development, including:

- First Generation LLM (2018-2019)
- Second Generation LLM (2020-2021)
- Advanced Conversational Capabilities (2022)
- Multimodal Capabilities (2022-2023)
- Autonomous Agent Capabilities (2023)
- Advanced Tool-Using Capabilities (2023-Present)
- Future Capabilities (Beyond 2023)

For each node, we provide concise descriptions, dangers, facts, capabilities, and subcategories, focusing on key points and bullet-style formatting for clarity.

We approached this project by:

- Data Compilation: Aggregating information on AI advancements, safety concerns, and economic impacts.
- Structured Representation: Organizing the information into a hierarchical tree data model for easy navigation and analysis.
- Safety Focus: Emphasizing the potential dangers and security challenges associated with each capability.
- Economic Analysis: Including facts about user adoption rates, market valuations, investment figures, and productivity impacts to highlight the economic significance.

While we cannot include actual screenshots here, the `treeData` structure creates a clear, visual representation of the AI capability evolution. Each node and subcategory can be visualized as part of an interactive timeline or tree diagram, allowing users to explore the information in a structured manner.

3. Code

Our codebase is structured around JavaScript, using JSON-like objects to represent the hierarchical tree data (`treeData`). The code includes:

- **Hierarchical Data Model:** The `treeData` object encapsulates all information, making it easily parsable and extendable.
- **Modular Design:** The structure allows for adding or modifying nodes without affecting the overall integrity.
- **Integration Capabilities:** The data model can be integrated into web applications, allowing for dynamic visualization (e.g., using D3.js or similar libraries).

The project references various large language models like GPT-2 and GPT-3, analyzing their capabilities and implications. While no specific AI models are implemented in the code, the project discusses the impact of these models on agent security. While the code primarily consists of data structures, it is designed to be compatible with visualization libraries for future expansion.

We ensured that all data included is accurate, avoiding disallowed content and respecting privacy concerns.

4. Discussion and Conclusion

Observations:

- **Rapid Progression:** AI capabilities have evolved rapidly, with significant advancements occurring within a few years. Each generation introduces more complex functionalities, raising new safety concerns.
- **Economic Impact:** AI advancements have substantial economic implications, influencing markets and productivity across industries.

The project's structured approach highlights how advancements can outpace safety considerations, emphasizing the need for proactive measures. By identifying potential dangers early, stakeholders can work towards aligning AI development with human values and safety requirements. The project underscores the importance of interdisciplinary approaches in AI safety, combining technical knowledge with ethical and economic perspectives. It can serve as a foundation for educators, policymakers, and researchers to understand and address the multifaceted challenges of AI agent security.

5. Appendix

Brynjolfsson, E., & Li, D. (n.d.). The Economics of Generative AI. Towards Data Science.

<https://towardsdatascience.com/economics-of-generative-ai-75f550288097>

Cognizant. (n.d.). Generative AI Economic Model.
<https://www.cognizant.com/us/en/aem-i/generative-ai-economic-model-oxford-economics#spy-DecodingJobDisruption>

Federal News Network. (2024, August). USACE data scientists enabling AI analytics across the Army.
<https://federalnewsnetwork.com/federal-insights/2024/08/usace-data-scientists-enabling-ai-analytics-across-army/>

Felicitis. (n.d.). The Agent Economy.
<https://www.felicitis.com/insight/the-agent-economy>

Fortune. (2024, May 23). U.S. intelligence agencies using generative AI 3 years before ChatGPT was released.
<https://fortune.com/2024/05/23/us-intelligence-agencies-using-generative-ai-3-years-chatgpt-was-released/>

Fowler, G. (n.d.). How Will Multi-Agent LLMs on Quantum Computers Change Everything? Medium.
<https://gafowler.medium.com/how-will-multi-agent-llms-on-quantum-computers-change-everything-bb640a2cdd71>

McKinsey. (n.d.). The Economic Potential of Generative AI.
<https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier>

TechStrong AI. (n.d.). Research Shows Risk in Using LLMs for Military Decision-Making.
<https://techstrong.ai/articles/research-shows-risk-in-using-llms-for-military-decision-making/>