Let's think step by step to build Autonomous Agents

Dina Adrian
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
adriangh.dina@gmail.com

Bogdan Ionescu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
bogdan.ionescu@upb.ro

***Abstract*—The emergence of autoregressive Large Language Models (LLMs) has enabled autonomous agents to surpass previous limitations. Early research in this field typically focused on training agents with restricted knowledge in isolated environments [1]. The learning capacities of transformer models [2], due to vast data volumes, demonstrate potential for achieving human-level intelligence, thus intensifying research into LLM-based autonomous agents [3]. This paper proposes a research and development strategy for agents in the context of LLMs.**

***Index Terms*—Agents, LLM, Reasoning, Planing, Memory, Action, Reinforcement Learning**

I. Introduction

In recent years, large-scale language models have achieved remarkable successes, indicating significant potential for human-like intelligence [4]. This capability arises from extensive training datasets combined with a large number of model parameters. Consequently, an increasingly important field has developed around using LLMs as central control systems to create autonomous agents capable of decision-making and action generation. LLM-based agents possess a broader internal understanding of the world, facilitating more informed actions, even without domain-specific training. Furthermore, LLM-based agents can provide natural language interfaces for human interaction, adding flexibility and explanatory capability.

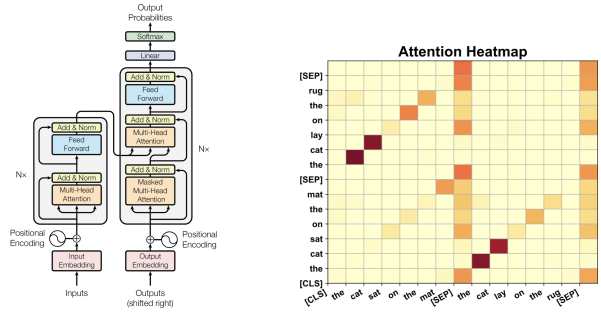
II. Foundation Models

A. Attention is All You Need

Transformer architectures offer a modern, innovative approach for efficiently managing continuous and large-scale data modeling tasks. The attention mechanism allows Transformers to capture relationships between sequence elements, regardless of their relative position. Distributed training across multiple GPUs enables the swift learning of large datasets [2].

B. Bidirectional Encoder Model

These models rely on Encoder blocks from the Transformer architecture, making them effective for tasks that require deep contextual and semantic understanding [5]. They are foundational models that can be fine-tuned for tasks such as sentiment analysis, entity recognition, and text classification.



(a) General architecture [2] (b) Attention Heatmap [8]

Fig. 1: Transformer model principal features

C. Sequence-to-Sequence Model

This model employs the full Transformer architecture, with an encoder for processing input text and a decoder for generating output text. T5 (Text-To-Text Transfer Transformer) treats all NLP tasks (e.g., summarization, text generation, machine translation) as text-to-text conversion problems [6].

D. Autoregressive Model

GPT (Generative Pretrained Transformer) is known as an autoregressive model for text generation. By masking values above the main diagonal in the attention mechanism matrices, these models are trained to predict the next token in a sequence based only on preceding tokens [7]. With extensive parameters, tasks like text classification or sentiment extraction become straightforward with Prompt Tuning techniques. These models are now widely used in developing Chat-bots or Virtual Agents that respond to users effectively and naturally.

III. Autonomous Agents

LLM-based autonomous agents are expected to perform various tasks effectively, leveraging human-like capabilities of these models. To achieve this, two essential aspects must be addressed: (1) selecting an architecture that optimally utilizes LLMs and (2) enabling the agent to develop task-specific skills once the architecture is defined [9]. Key capabilities of an autonomous agent include reasoning, memory, planning, and action execution.

1st Workshop on Artificial Intelligence for Multimedia, AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest, Bucharest, Romania, 8 November, 2024

A. Reasoning

Reasoning is the ability to process information, establish conceptual links, and draw logical conclusions based on accumulated knowledge and experiences. For virtual and autonomous agents, reasoning is essential for making intelligent decisions, understanding and interpreting complex situations, and adapting responses to environmental challenges.

At first, reasoning capabilities were highlighted through Prompt Engineering techniques. For instance, a simple prompt such as "Let's think step-by-step ..." [10] can enhance the generation performance of a valid response more effectively than applying fine-tuning techniques to the model. GPT-4 employs Reinforcement Learning techniques to generate complex responses and actions, enabling it to produce a long internal Chain of Thought before arriving at a final response.

B. Memory

Memory in autonomous agents, particularly LLM-based agents, refers to the agent's ability to store and manage information obtained during interactions for later use. This enables learning from experience, maintaining conversational context, and providing more personalized and relevant interactions [11].

C. Planing

Planning is the capability of an autonomous agent to organize a sequence of steps necessary to achieve a specific goal. In LLM-based agents, planning is supported by the model's ability to understand and analyze complex contexts and generate logical action sequences [9].

D. Action

In the context of autonomous agents, action refers to the agent's capability to perform specific tasks or operations in response to its environment or assigned tasks. This is the final stage in decision-making, involving the translation of plans and strategic decisions into observable behaviors and tangible outcomes [12].

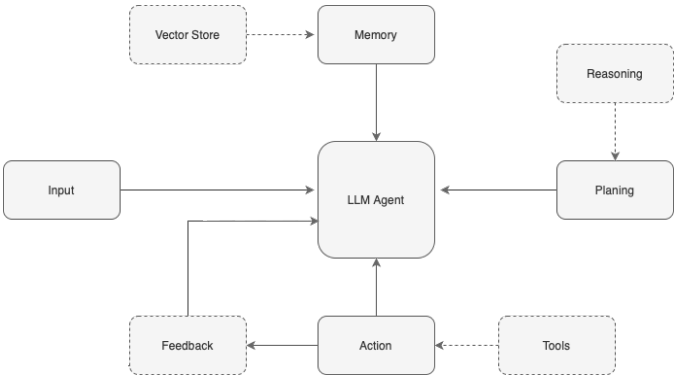


Fig. 2: Example of Agent Architecture

IV. Conclusions and Future Development

Virtual agents are an expanding field, significantly impacting how we interact with technology. From virtual assistants facilitating daily activities to autonomous agents performing complex tasks across various domains, these intelligent entities have shown great potential for improving efficiency and delivering innovative solutions. Future research areas in the development of autonomous virtual agents include:

- Explainability to enhance reasoning capabilities in text.
- Use of expert systems for transformer-based networks (MoE).
- Application of Reinforcement Learning and its derivatives (RLHF, DPO, PPO).
- Utilization of multimodal models for integrating understanding across various types of information (e.g., image, audio).
- Efficient training and fine-tuning techniques, such as Parameter Efficient Fine-Tuning (PEFT), Low Rank Adaptation (LoRA) and Differential Transformers.
- Prompt-Tuning and decision-making enhancements based on context.
- Evaluation over multiple agent benchmark.
- Proposal for a new Agent architecture.

References

[1] L.-B. K. Shoham Y, *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[3] F. C. Stefano V. Albrecht and L. Schäfer, *Multi-Agent Reinforcement Learning Foundations and Modern Approaches*. The MIT Press Cambridge, Massachusetts London, England, 2024.

[4] OpenAI, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2023. [Online]. Available: https://arxiv.org/abs/1910.10683

[7] M. Hanna, O. Liu, and A. Variengien, "How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model," 2023. [Online]. Available: https://arxiv.org/abs/2305.00586

[8] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," 2023. [Online]. Available: https://arxiv.org/abs/2309.01029

[9] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, Mar. 2024. [Online]. Available: http://dx.doi.org/10.1007/s11704-024-40231-1

[10] M. Jin, Q. Yu, D. Shu, H. Zhao, W. Hua, Y. Meng, Y. Zhang, and M. Du, "The impact of reasoning step length on large language models," 2024. [Online]. Available: https://arxiv.org/abs/2401.04925

[11] Z. Zhang, X. Bo, C. Ma, R. Li, X. Chen, Q. Dai, J. Zhu, Z. Dong, and J.-R. Wen, "A survey on the memory mechanism of large language model based agents," 2024. [Online]. Available: https://arxiv.org/abs/2404.13501

[12] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," 2023. [Online]. Available: https://arxiv.org/abs/2302.04761