UNIVERSITY OF TRENTO

Department of Information Engineering and Computer Science

Master's Degree in
Artificial Intelligence Systems

Final Dissertation

Exploring the Use of LLMs for Agent Planning: Strengths and Weaknesses

Supervisor
Paolo Giorgini

Student
Davide Modolo
229297

Academic year 2023/2024

Contents

Abstract

1 Introduction

2 Background

In this thesis, we will analyze in detail the behavior of an LLM as an agent within a controlled environment.

Before presenting all the work carried out in detail, this chapter aims to provide a comprehensive explanation of all the theoretical foundations necessary to understand the steps presented in the following chapters. Starting from a brief introduction of Artificial Intelligence just to define the boundaries in which we are working, we will move to the core concepts. In particular, we want to highlight what an LLM is and how it works, with a special focus on the Attention mechanism and how the uncertainty of an LLM can be calculated. This will serve as a basis for correctly interpreting the results analyzed in Section 6.

There will also be a broader discussion on agents in a strict sense and "LLM agents" to better show the difference between our implementation and what is currently being discussed over the media.

To better define the context of this thesis, we will examine the main alternative approaches to solving a logistical problem currently studied in the literature.

2.1 Artificial Intelligence

Right now in the media, AI is being used as a synonym for Large Language Models. However, AI is a broader concept that includes many techniques and methodologies.

Summary of different kind of AI ending with Generative Models end with NLP history

2.2 Large Language Models - LLMs

LLMs are generative models released with the paper "Attention is All You Need"

2.2.1 LLMs' Uncertainty

Understanding the uncertainty of an LLM is crucial to correctly interpret the text it generates. If we ask for a yes/no question, it would make a different impact on us reading "Yes" or "Yes - Uncertainty 49%". Moreover, this would let us get some kind of explainability behind these complex and opaque systems.

cite Hallucinations

2.3 Agents

As widely explained in the book "An Introduction to Multiagent Systems" [2], we can summarize the definition of an agent as an autonomous entity that perceives its environment through sensors and acts upon it through effectors, making decisions based on its perceptions and objectives in order to achieve specific goals.

This definition highlights several key aspects of agents:

- Autonomy: Agents operate without direct human intervention, controlling their own actions.
- Perception and Action: They interact with the environment via sensors (perception) and effectors (action execution).
- Decision-making: Agents select actions based on their internal model, goals, and the state of the environment.
- Non-determinism and Adaptability: Since environments are generally non-deterministic, agents must be prepared for uncertainty and potential failures in action execution.

- Preconditions and Constraints: Actions are subject to certain conditions that must be met for successful execution.

Thus, an agent's fundamental challenge is deciding which actions to perform in order to best satisfy its objectives, given the constraints and uncertainties of its environment.

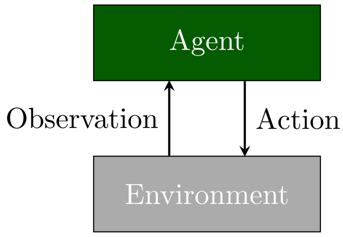

Figure 2.1: Agent Scheme
Source: redesign of a scheme in [2]

As shown in Figure 2.1, an agent is some entity that perceives the environment and reacts to it. The environment can be anything from a simple thermostat to a complex system like a self-driving car. The idea is that the agent is able to react to a change in the environment and take actions to achieve its goals.

We will analyze in detail the prompts and the choices in the Chapter 5 Section 5.2, but to give an some kind of help to align with the definition above, we can map some of its concept to what this thesis will analyze:

- Perception and Action: what the server sends about the current state of the environment can be seen as the perception of the agent, while the action it can take will be given in the prompt in a specific way.

- Decision-making: the decision-making process will be the generation of the text by the LLM, weighted by the uncertainty.

- Non-determinism and Adaptability: to emulate the non-determinism of the environment, the state received by the server will be used "raw" in the prompt, without any hard processing or parsing.

- Preconditions and Constraints: living in a "limited" map with a fixed number of cells, is itself a constraint the agent must consider.

2.3.1 BDI Architecture

The Belief-Desire-Intention (BDI) architecture is a widely adopted framework in artificial intelligence (AI) for modeling rational agents. It was formally developed by Rao and Georgeff in 1995 [1] and has been implemented in several architectures, including PRS (1987), dMARS (1998), JAM (1999), Jack (2001), and JADEX (2005). BDI provides a structured approach to practical reasoning, allowing agents to function effectively in dynamic and unpredictable environments.

2.3.1.1 Core Components of BDI

BDI agents operate based on three key components:

- Belief: Represents the agent's knowledge about the world, including past events and observations. Given the agent's local perception and limited computational resources, beliefs act as cached, imperfect information rather than absolute knowledge

- Desire (Goals): Defines the agent's objectives or preferred end states, such as "desiring to graduate." Desires help in justifying why an agent takes specific actions and enable reasoning about goal interactions, particularly in failure recovery scenarios

- Intention: Represents the commitments of an agent toward achieving specific goals through selected plans. Intentions provide structure by ensuring persistence and internal consistency, allowing for incremental planning and adaptation in real-time. They also serve as a means of coordinating multiple agents in distributed systems.

2.3.1.2 BDI in Practical AI Applications

BDI has been extensively used in fields like robotics, automated planning, and multi-agent systems. For instance, JADEX is a popular Java-based BDI framework used in simulations and decision-making applications.

2.4 State of the Art

A logistic problem is a typical problem in the field of AI, and there are several approaches to solve it. In this section, we will analyze the main alternatives to LLMs in the literature and highlight the core differences between them and our approach.

> This is a longer blockquote that may span multiple paragraphs. It is indented on both sides.

> This is another styled blockquote with a grey background.

2.4.1 PDDL Based Solution

2.4.1.1 Planning Domain Definition Language

PDDL: pro (explainable, efficienza in ambienti limitati) e contro (time intensive, non c'è adattamento, in un problema di dimensioni moderate è quasi impossibile usarlo in real time)

Reinforcement Learning: pro (efficace, si può adattare all'ambiente che cambia) e contro (convergenza ad un local minima nella reward function, no explainability, costoso da allenare)

Agenti LLM (es. o1 di OpenAI): introduzione a temi emergenti come il Chain of Thought, sottolineando come viene eseguito il reasoning e cosa differisce dal mio approccio

Planning in LLM

3 Experiment Setting

3.1 Problem Definition

3.2 Environment - Deliveroo.js

3.3 GPT Models

4 Agent Development

4.1 First Approach

4.2 Second Approach

4.3 Final Agent

4.4 Closest Cell to the Goal

5 Data Collection

5.1 Visualize the Attention

5.2 Prompts

5.3 Prompt Creation Choices

5.4 Heatmap Generation

6 Results Discussion

6.1 Stateless

6.2 Stateful

6.3 Stateless and Stateful Combined results

6.4 Closest Cell to the Goal Problems

6.5 Models Comparison

7 Future Works

8 Conclusions

Bibliography

[1] Anand S. Rao and Michael P. Georgeff. BDI agents: From theory to practice. In Victor R. Lesser and Les Gasser, editors, *1st International Conference on Multi Agent Systems (ICMAS 1995)*, pages 312–319, San Francisco, CA, USA, 12-14 June 1995. The MIT Press.

[2] Michael Wooldridge. *An Introduction to Multiagent Systems.* John Wiley & Sons, 1 edition, 2002. Chapter 2.

Appendix A Attachment

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.