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

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

2.3.1 BDI Architecture

2.4 State of the Art

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

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.