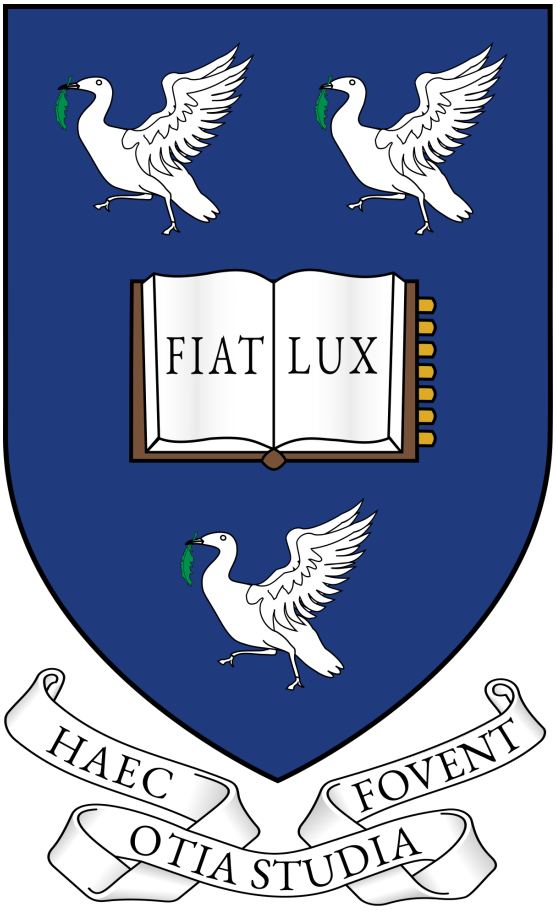

THE UNIVERSITY OF LIVERPOOL

COMP702-MSC FINAL PROJECT

Visual Spatial Reasoning for Large Language Models

Author Jinlong Liu
ID 201678181
Supervisor Frank Wolter

June 21, 2023

I. Project Description

In recent years, Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP) with their remarkable achievements. These models, trained on vast amounts of text data, have proved highly effective in solving numerous NLP tasks, including machine translation, question answering, and text generation. Their ability to comprehend and generate human-like text has been nothing short of impressive.

However, despite their proficiency in various linguistic tasks, LLMs have encountered challenges when it comes to Spatial Reasoning. Unlike tasks centered purely on language, Spatial Reasoning requires understanding and manipulating visual and spatial information. To shed light on this limitation, I aim to conduct an experiment utilizing popular LLMs like GPT-4. The focus will be on testing their capability in Visual Spatial Reasoning.

The experiment will revolve around a widely used task known as Visual Question Answering (VQA), which demands machines to answer questions based on provided images. For instance, given an image depicting different objects, the machine will be tasked with responding to inquiries such as "What is the spatial relationship between object A and object B?". Through these carefully crafted tests, we can gauge the extent of Spatial Reasoning proficiency exhibited by LLMs.

By assessing their performance in VQA and their ability to comprehend and reason about spatial relationships in visual content, we can obtain valuable insights into the Spatial Reasoning capabilities of LLMs. These findings will not only deepen our understanding of the strengths and limitations of these models but also pave the way for further advancements in the realm of NLP, bringing us closer to more comprehensive and versatile language models

II. Aims and Objectives

A. Aims

1) To investigate the Spatial Reasoning abilities of Large Language Models (LLMs), specifically focusing on their performance in Visual Question Answering (VQA) tasks.
2) To understand the limitations and challenges faced by LLMs in comprehending and reasoning about spatial relationships in visual content.
3) To assess the current state of Spatial Reasoning capabilities in leading LLMs, then compare with elder models and explore their improvement in this domain. Besides, to identify potential avenues for enhancing LLMs' Spatial Reasoning abilities.

B. Objectives

1) Design and develop a comprehensive experimental framework for evaluating LLMs' Spatial Reasoning abilities in VQA tasks.
2) Curate a diverse dataset of visual stimuli and corresponding questions that require spatial reasoning skills to answer accurately.
3) Conduct systematic experiments using the prepared dataset and modified LLMs to assess their performance and measure their level of Spatial Reasoning competence.
4) Analyze the experimental results to identify patterns, trends, and challenges in LLMs' Spatial Reasoning capabilities.
5) Provide insights into the strengths and limitations of current LLMs for Spatial Reasoning tasks, based on the experimental findings.
6) Suggest potential avenues for enhancing LLMs' Spatial Reasoning abilities, such as incorporating multimodal information or novel architectural modifications.
7) Contribute to the existing body of knowledge in the field of NLP by advancing our understanding of LLMs' Spatial Reasoning abilities and their implications for future research and development.

III. Key Literature and Background Reading

IV. Development and Implementation Summary

V. User Interface Mockup

VI. Data Sources

VII. Testing

VIII. Evaluation

IX. Ethical Considerations

X. Project Plan

XI. Risks and Contingency Plans

References

[1] A. Borji, "A categorical archive of chatgpt failures," *arXiv preprint arXiv:2302.03494*, 2023.