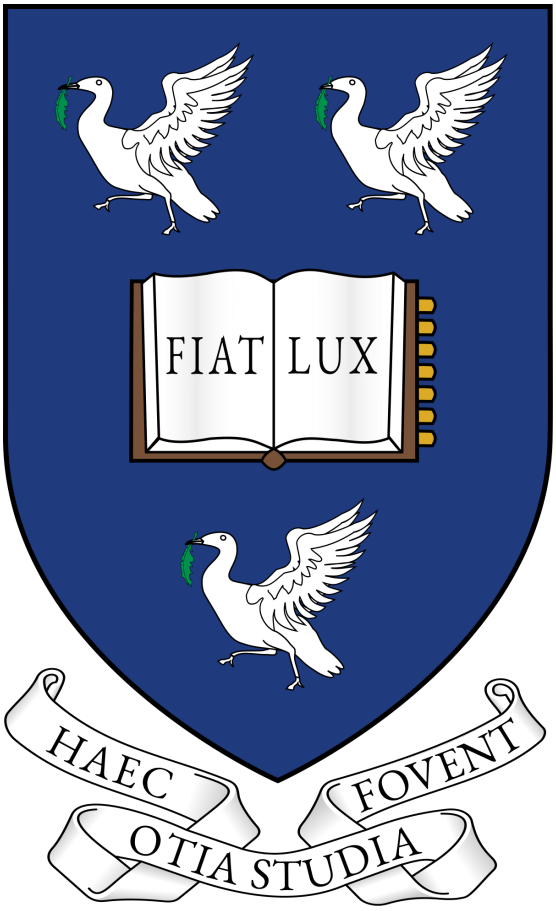


THE UNIVERSITY OF LIVERPOOL

COMP702-MSC FINAL PROJECT

# Visual Spatial Reasoning for Large Language Models

*Author* Jinlong Liu
*ID* 201678181
*Supervisor* Frank Wolter

**June 24, 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 [1]. 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

In this section, we will provide a clear outline of the aims and objectives of this project. The aims represent the overarching goals that we aim to achieve, while the objectives outline the specific steps and milestones that will be undertaken to fulfill these goals. The aims and objectives of this project are outlined below:

### 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

Spatial Reasoning ability is a crucial aspect of human intelligence, enabling us to understand and manipulate spatial information. It is a fundamental skill that allows us to navigate the world around us, comprehend visual content, and solve spatial problems. It is also a goal to be achieved by Artificial Intelligence (AI) systems, as it is a key component of human-level intelligence. At first, the representation of spatial knowledge having a different way of thinking from the representation of language knowledge, it is difficult to combine the two. However, with the development of deep learning, the combination of the two has become possible.

For Large Language Models (LLMs), Spatial Reasoning encompasses several key aspects, including (i) mereotopology, (ii) direction and orientation, (iii) size, (iv) distance, and (v) shape, as discussed by Cohn et al. (2008) in their qualitative analysis [2]. To assess LLMs' performance in Spatial Reasoning, Cohn et al. (2023) developed a comprehensive Spatial Reasoning test that incorporates basic spatial relations (such as parthood, rotation, and direction), size, shape, location, affordances, object interaction, and object permanence [3]. Their findings revealed inconsistencies in LLMs' responses to the same question during continuous sessions, indicating that LLMs struggle with maintaining consistency in Spatial Reasoning tasks.

In a related study by Lin et al. (2023), a test was devised to evaluate LLMs' Spatial Reasoning abilities with a focus on mathematical representation of questions [4]. Their research demonstrated that while LLMs may not excel at directly answering spatial questions, they exhibit proficiency in transforming questions into mathematical forms based on user demands.

Moreover, Borji (2023) proposed a failure task category for ChatGPT, within which Spatial Reasoning was highlighted [5]. In this task, the researcher provided a grid description and asked the model to find a path from a start point to an end point. However, the model failed to answer the question accurately, highlighting its limitations in Spatial Reasoning tasks.

Collectively, these studies shed light on the challenges LLMs face when it comes to Spatial Reasoning, including inconsistencies in responses, struggles with direct spatial question answering, and difficulties in solving specific spatial tasks like path finding.

One important aspect to consider is Visual Spatial Reasoning (VSR) and its relevance in the context of Visual Language Models (VLMs), which is a distinct topic from LLMs. It is worth highlighting some related research in this field.

Researchers have dedicated efforts to assess the ability of VLMs in comprehending commonsense spatial reasoning. Visual Question Answering (VQA) tasks have been employed to evaluate the performance of VLMs in answering questions that involve spatial relationships between objects depicted in images [6]. Consequently, various benchmarks and question sets have been developed to specifically evaluate the spatial reasoning capabilities of VLMs.

For instance, Liu et al. (2022) constructed a dataset consisting of over 10,000 text-image pairs, encompassing 66 types of spatial relations [7]. They evaluated the performance of both human subjects and VLMs on this dataset, revealing that VLMs still lag behind humans in this task. Additionally, other benchmarks such as GQA [8], AGQA [9], Compositional Visual Relations Dataset [10], VALSE [11], Liu et al. (2021) [12], and SpartQA [13] have been developed to further gauge the spatial reasoning capabilities of VLMs.

These benchmarks and datasets serve as valuable tools to assess the progress and limitations of VLMs in the realm of spatial reasoning and they have such things in common: (i) they are all based on commonsense spatial reasoning, (ii) they all focus on VLMs, and (iii) they all given some bound words to the model to help it understand the spatial relations. However, in my research, I will focus on LLMs and their Spatial Reasoning capabilities, which is a distinct topic from VLMs.

## IV. Development and Implementation Summary

This project aims to investigate the Visual Spatial Reasoning capabilities of LLMs, specifically focusing on their performance in VQA tasks. So the development and implementation of this project will be divided into these parts:

### A. Probelm Identification

The main problem of this project is to investigate the Spatial Reasoning abilities of LLMs, specifically focusing on their performance in VQA tasks. So the first step is to identify the problem and the research questions. The research questions are as follows:

1) What are the Spatial Reasoning abilities of LLMs?
2) What are the limitations and challenges faced by LLMs in comprehending and reasoning about spatial relationships in visual content?
3) What is the current state of Spatial Reasoning capabilities in leading LLMs?
4) What are the potential avenues for enhancing LLMs' Spatial Reasoning abilities?

### B. Experiment Design

For the experimental stage, the workflow can be visualized as depicted in Figure 1. The experiment is divided into three distinct parts, each serving a specific purpose:

1) **Prepare Stage**
   During this stage, the primary objective is to gather a diverse collection of images that contain various spatial relations. These images can be sourced from the internet or generated using appropriate tools. For this project, the dataset will consist of a combination of an existing benchmark proposed by Liu et al. (2022) [7]. Another crucial aspect is the question set, which will be divided into two parts: (i) Questions discussing the spatial relationships between objects depicted in the images. (ii) Questions requiring the reasoning steps behind the previous answers.
2) **Experiment Stage**
   The focus of this stage is to test different versions of Large Language Models (LLMs) using the prepared dataset and question set. The LLMs will be evaluated based on two factors: (i) Different sizes of LLMs. (ii) Different pre-training methods of LLMs. The experiment will involve running the LLMs on the prepared dataset, recording the results, and analyzing their performance. And all testing models will be GPT models, including from GPT-3.5 tubor to GPT-4-0612.
3) **Analysis and Evaluation Stage**
   In this stage, the emphasis is on analyzing the performance of the different versions of LLMs using the prepared dataset. The analysis will primarily consider two aspects: (i) The accuracy of the answers provided by the LLMs. (ii) The reasoning steps generated by the LLMs. Based on the accuracy, the effectiveness and capability of the LLMs in spatial reasoning can be evaluated and compared. By examining the reasoning steps, any limitations and challenges faced by LLMs in comprehending and reasoning about spatial relationships in visual content can be identified.

Through this systematic workflow, the experiment aims to assess the performance and capabilities of various versions of LLMs in spatial reasoning tasks. The analysis and evaluation stage provide insights into the strengths, weaknesses, and potential areas for improvement in LLMs' understanding and reasoning abilities pertaining to spatial relationships in visual content.

### C. Code Implementation

There is no code implementation in this project. All the experiments will be conducted on the website of OpenAI. The

code of this project will be written in Python, and the code will be used to prepare the dataset and question set, and to analyze the results of the experiment.

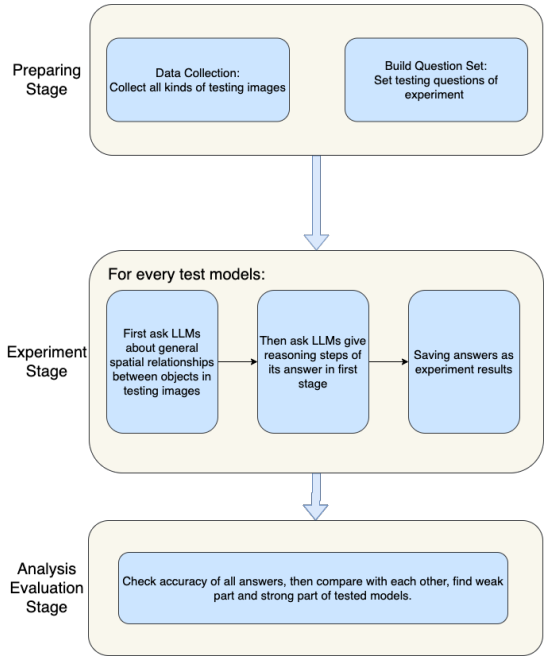


Fig. 1. Workflow of the experiment

## V. User Interface Mockup

For this project, there is no user interface mockup. All the experiments will be conducted on the website of OpenAI. The code of this project will be written in Python, and the code will be used to prepare the dataset and question set, and to analyze the results of the experiment.

## VI. Data Sources

The data utilized in this project will be divided into two main parts: the dataset and question set, and the answers generated by the Large Language Models (LLMs).

For the dataset and question set, a combination of existing benchmarks proposed by Liu et al. (2022) [7] and self-generated images will be used. The images in the dataset will encompass real-world objects along with 2D shapes, ensuring a diverse range of spatial relationships to be examined. The benchmark images are extracted from the COCO 2017 dataset [14], which is an open-source dataset widely used in computer vision research. It is important to note that the usage of the COCO dataset has been granted full permission for this project.

As for the second part, the answers generated by the LLMs will be stored in a text file. The answers will be generated using the OpenAI website, which provides a platform for executing queries and obtaining responses from the language model. The generated answers can be conveniently downloaded as a text file from the website or saved directly on the platform.

By utilizing this comprehensive data setup, incorporating a combination of benchmark images and self-generated content, along with leveraging the power of LLMs to generate answers, the project aims to provide an extensive and diverse set of data for analysis and evaluation. The combination of the benchmark dataset and the generated answers will enable the project to assess the performance and capabilities of the LLMs in spatial reasoning tasks accurately.

It is important to ensure adherence to proper data usage policies, including obtaining necessary permissions for the benchmark dataset and complying with any licensing or attribution requirements associated with the dataset. This approach ensures the project's integrity and ethical practices while utilizing publicly available data resources for research purposes.

## VII. Testing

## VIII. Evaluation

## IX. Ethical Considerations

## X. Project Plan

The project timeline and task allocation are visually represented in Figure 2, showcasing the specific Gantt chart. The project is organized into four distinct stages, each comprising multiple tasks aimed at achieving the project objectives. The stages and their corresponding tasks are outlined below:

1) Literature Review
   this
   •
2) Build Dataset
3) Experiment Stage
4) Analysis and Evaluation Stage
5) Final Report

## XI. Risks and Contingency Plans

## References

[1] B. Peng, C. Li, P. He, M. Galley, and J. Gao, "Instruction tuning with gpt-4," *arXiv preprint arXiv:2304.03277*, 2023.
[2] A. G. Cohn and J. Renz, "Qualitative spatial representation and reasoning," *Foundations of Artificial Intelligence*, vol. 3, pp. 551–596, 2008.
[3] A. G. Cohn and J. Hernandez-Orallo, "Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of llms," *arXiv preprint arXiv:2304.11164*, 2023.
[4] F. Lin, Z. Shou, and C. Chen, "Using language models for knowledge acquisition in natural language reasoning problems," *arXiv preprint arXiv:2304.01771*, 2023.
[5] A. Borji, "A categorical archive of chatgpt failures," *arXiv preprint arXiv:2302.03494*, 2023.
[6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2425–2433.
[7] F. Liu, G. Emerson, and N. Collier, "Visual spatial reasoning," *arXiv preprint arXiv:2205.00363*, 2022.
[8] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.
[9] M. Grunde-McLaughlin, R. Krishna, and M. Agrawala, "Agqa: A benchmark for compositional spatio-temporal reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 287–11 297.

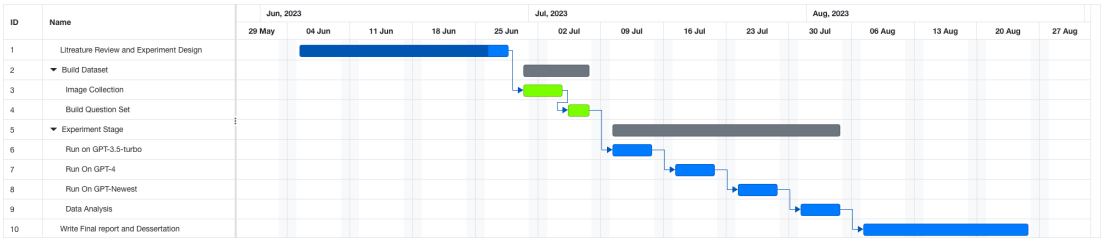


Fig. 2. Gantt Chart

[10] A. Zerroug, M. Vaishnav, J. Colin, S. Musslick, and T. Serre, "A benchmark for compositional visual reasoning," *arXiv preprint arXiv:2206.05379*, 2022.

[11] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, and A. Gatt, "Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena," *arXiv preprint arXiv:2112.07566*, 2021.

[12] X. Liu, D. Yin, Y. Feng, and D. Zhao, "Things not written in text: Exploring spatial commonsense from visual signals," *arXiv preprint arXiv:2203.08075*, 2022.

[13] R. Mirzaee, H. R. Faghihi, Q. Ning, and P. Kordjmashidi, "Spartqa:: A textual question answering benchmark for spatial reasoning," *arXiv preprint arXiv:2104.05832*, 2021.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.